DOCUMENT RESUME

ED 403 326 TM 026 072

AUTHOR	Crocker, Linda; Zieky, Michael
TITLE	Joint Conference on Standard Setting for Large-Scale Assessments (Washington, D.C., October 5-7, 1994). Proceedings, Volume II.
INSTITUTION	Aspen Systems Corp., Rockville, MD.
SPONS AGENCY	National Assessment Governing Board, Washington, DC.; National Center for Education Statistics (ED), Washington, DC.
PUB DATE	Oct 95
NOTE	436p.; For Volume I, the executive summary, see TM 026 071.
AVAILABLE FROM	U.S. Government Printing Office, Superintendent of Documents, Mail Stop SSOP, Washington, DC 20402-9328.
PUB TYPE	Collected Works - Conference Proceedings (021)
EDRS PRICE	MF01/PC18 Plus Postage.
DESCRIPTORS	*Academic Achievement; *Accountability; *Educational Assessment; Educational Improvement; Educational Policy; *Elementary Secondary Education; Program Evaluation; *Standards; Statistical Bias; Testing Problems; Test Use; Validity
IDENTIFIERS	*Large Scale Assessment; *Standard Setting

ABSTRACT

The National Assessment Governing Board and the National Center for Education Statistics sponsored a Joint Conference on Standard Setting for Large-Scale Assessments to provide a forum for technical and policy issues relevant to setting standards at local, state, and national levels. Volume I contains an executive summary of the conference and synopses of the conference papers. This volume comprises the papers prepared for the conference and summaries of the plenary sessions and small breakout sessions. Educators were invited to present papers on issues within historical, theoretical, methodological, application, or policy perspectives. The 19 invited papers explore 6 major themes identified by the authors of the executive summary. These are: (1) multiple meanings and uses of standards; (2) methods of setting standards; (3) new directions and technical issues in setting standards; (4) fairness and validity in setting standards; (5) problems and controversies; and (6) areas of agreement in setting and using standards. While the conference did not result in consensus on how standards ought to be set for large scale assessments, it did bring together many experts in the field, and it did promote an understanding of the multifaceted issues involved in standard setting. Each paper contains references. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 403 326

THE NATION'S
REPORT
CARD
naep

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

JOINT CONFERENCE ON STANDARD SETTING FOR LARGE-SCALE ASSESSMENTS

PROCEEDINGS

VOLUME II

TM026072

Prepared by Aspen Systems under contract with the
National Assessment Governing Board and the National
Center for Education Statistics

BEST COPY AVAILABLE

What is The Nation's Report Card?

THE NATION'S REPORT CARD, the National Assessment of Educational Progress (NAEP), is the only nationally representative and continuing assessment of what America's students know and can do in various subject areas. Since 1969, assessments have been conducted periodically in reading, mathematics, science, writing, history/geography, and other fields. By making objective information on student performance available to policymakers at the national, state, and local levels, NAEP is an integral part of our nation's evaluation of the condition and progress of education. Only information related to academic achievement is collected under this program. NAEP guarantees the privacy of individual students and their families.

NAEP is a congressionally mandated project of the National Center for Educational Statistics, the U.S. Department of Education. The Commissioner of Education Statistics is responsible, by law, for carrying out the NAEP project through competitive awards to qualified organizations. NAEP reports directly to the Commissioner, who is also responsible for providing continual reviews, including validation studies and solicitation of public comment, on NAEP's conduct and usefulness.

The National Assessment Governing Board (NAGB) is established under section 412 of the National Education Statistics Act of 1994 (Title IV of the Improving America's Schools Act of 1994, Pub. L. 103-382). The Board is established to formulate policy guidelines for the National Assessment of Educational Progress. The Board is responsible for selecting subject areas to be assessed, developing assessment objectives, identifying appropriate achievement goals for each grade and subject tested, and establishing standards and procedures for interstate and national comparisons.

The National Assessment Governing Board

Honorable William T. Randall, Chairman
Commissioner of Education
State Department of Education
Denver, Colorado

Mary R. Blanton, Vice-Chair
Attorney
Blanton & Blanton
Salisbury, North Carolina

Honorable Evan Bayh
Governor of Indiana
Indianapolis, Indiana

Patsy Cavazos
Principal
W.G. Love-Accelerated Elem. School
Houston, Texas

Honorable Naomi K. Cohen
Director of Purchased Services
Planning and Policy Division
Hartford, Connecticut

Charlotte Crabtree
Professor of Education
University of California
Los Angeles, California

Catherine L. Davidson
Secondary Education Director
Central Kitsan School District
Silverdale, Washington

James E. Ellingson
Fourth-Grade Classroom Teacher
Probstfield Elementary School
Moorhead, Minnesota

Chester E. Finn, Jr.
John M. Olin Fellow
Hudson Institute
Washington, D.C.

Michael J. Guerra
Executive Director
National Catholic Education
Association
Washington, D.C.

William J. Hume
Chairman of the Board
Basic American, Inc.
San Francisco, California

Jan B. Loveless
District Communications Specialist
Midland Public Schools
Midland, Michigan

Marilyn McConachie
Member, Board of Education
Glenbrook High Schools
Glenview, Illinois

Honorable Stephen E. Merrill
Governor of New Hampshire
Concord, New Hampshire

Jason Millman
Professor of Educational Research
Methodology
Cornell University
Ithaca, New York

Honorable Richard P. Mills
Commissioner of Education
State Department of Education
Montpelier, Vermont

William J. Moloney
Superintendent of Schools
Calvert County Public Schools
Prince Frederick, Maryland

Mark D. Musick
President
Southern Regional Education Board
Atlanta, Georgia

Mitsugi Nakashima
President
Hawaii State Board of Education
Honolulu, Hawaii

Michael T. Nettles
Professor of Education and Public Policy
University of Michigan
Ann Arbor, Michigan

Edgar D. Ross
Attorney
Fredriksted, St. Croix
U.S. Virgin Islands

Fannie N. Simmons
Math Specialist
Midlands Improving Math & Science Hub
Columbia, South Carolina

Marilyn A. Whirry
English Teacher
Mira Costa High School
Manhattan Beach, California

Sharon P. Robinson (Ex-Officio)
Assistant Secretary
U.S. Department of Education
Washington, D.C.

Roy Truby
Executive Director
NAGB
Washington, D.C.

PROCEEDINGS OF THE JOINT CONFERENCE ON STANDARD SETTING FOR LARGE-SCALE ASSESSMENTS OF THE NATIONAL ASSESSMENT GOVERNING BOARD (NAGB) AND THE NATIONAL CENTER FOR EDUCATION STATISTICS (NCES)

Hyatt Regency-Capitol Hill
400 New Jersey Avenue, NW
Washington, DC

October 5-7, 1994

Perspectives on Standard Setting for Large-Scale Assessments

• **HISTORICAL**	Describes what the U.S. and other countries learned in 40 years of education standard setting.
• **THEORETICAL**	Examines approaches to conceptualizing standard setting.
• **METHODOLOGICAL**	Explores alternative models for setting standards in education.
• **APPLICATIONS**	Considers technical qualities that yield appropriate interpretations of data.
• **POLICY**	Examines the policy impact of standards in education.

October 1995

Prepared by Aspen Systems Corporation under contract with the National Assessment Governing Board and the National Center for Education Statistics

National Assessment Governing Board
Roy Truby
Executive Director

National Center for Education Statistics
Emerson J. Elliott
Commissioner

Conference Planning Committee

Mary Lyn Bourque National Assessment Governing Board	***Peggy Carr*** National Center for Education Statistics
Sharif Shakrani National Center for Education Statistics	***Daniel Taylor*** National Assessment Governing Board

For ordering information on these proceedings, write to:

U.S. Government Printing Office
Superintendent of Documents
Mail Stop SSOP
Washington, DC 20402-9328

The contents of this publication resulted from a conference sponsored jointly by the National Assessment Governing Board and the National Center for Education Statistics of the United States Department of Education. The opinions, interpretations, and conclusions of the authors are their own and do not necessarily represent those of the sponsors.

Foreword

The National Assessment Governing Board (NAGB) and the National Center for Education Statistics (NCES) sponsored this Joint Conference on Standard Setting for Large-Scale Assessments in October 1994. The primary purpose of the conference was to provide a forum to address technical and policy issues relevant to setting standards for large-scale educational assessments at the national, state, and local levels.

Nationally and internationally renowned educators were invited to present papers on specific issues within each of the five perspectives:

1. *Historical*: what the United States and other countries have learned in 40 years of educational standard setting;
2. *Theoretical*: approaches to conceptualizing standard setting;
3. *Methodological*: alternative models for setting performance standards in education;
4. *Application*: relevant technical qualities that yield appropriate interpretations of standard setting data (generalizability, validity, fairness, and clarity of communication); and
5. *Policy*: the policy impact of standard setting at the national, state, and local levels, including the use of standards in constructing large-scale assessments.

The presenters participated in small group sessions that focused on each topic. Interactions and discussions between presenters and participants from throughout the nation added to the quality of information gathered and reported in these Proceedings.

In examining various issues surrounding setting performance standards on educational assessments, the Joint Conference achieved the following goals:

- Established some theoretical and empirical foundations for conceptualizing and designing performance standards in large-scale assessments;
- Identified different areas of concern regarding performance standards for large-scale assessments;
- Provided NAGB and NCES with guidance in examining alternative methodologies for developing standards; and
- Informed NAGB and NCES about relevant issues related to setting student achievement levels for the various subjects included in the National Assessment of Educational Progress (NAEP).

Volume I of the conference Proceedings includes an executive summary of the conference and synopses of the conference papers. Volume II comprises the papers prepared for the conference and summaries of the plenary sessions and small breakout sessions. The conference Proceedings provide a rich and valuable source of information for standard setting and should be of great interest to educators and policymakers.

Acknowledgements

Planning and execution of the Joint Conference on Standard Setting for Large-Scale Assessments involved the participation of numerous staff from the National Assessment Governing Board (NAGB), the National Center for Education Statistics (NCES), Educational Testing Service (ETS), American College Testing (ACT), and Aspen Systems Corporation, the logistical services contractor.

The Planning Committee comprising Mary Lyn Bourque and Daniel Taylor from NAGB and Peggy Carr and Sharif Shakrani from NCES provided key guidance for all phases of conference organization and execution. Arnold Goldstein of NCES worked closely with the Planning Committee in organizing the panels, working with the authors, and coordinating the publication process. Patricia Dabbs and Lawrence Ogle of NCES, and Mary Ann Wilmer of NAGB provided invaluable support in the organization of the conference. Summaries of the break-out sessions were prepared by Ruth Palmer. Lilly Gardner, Munira Mwalimu, Darice Stephenson, and Juanita Taylor from Aspen Systems Corporation provided logistical support for the conference and supported publication of the conference Proceedings.

First, we would like to thank the authors and discussants whose work appears in the two volumes of these Proceedings. Their informative presentations and learned papers provide a collection of some of the best current thinking about setting standards for large-scale performance assessments.

Special thanks also go to James Popham of IOX Assessment Associates who not only moderated a session and summarized the findings from the break-out sessions, but provided insight into standard setting based on his long experience in the area of assessment. We would also like to thank the conference moderators: Joseph Conaty, Office of Educational Research and Improvement; Michael Feuer, National Academy of Sciences; Sylvia Johnson, Howard University; and Michael Nettles, University of Michigan.

The facilitators and recorders of the small-group sessions contributed to a large part of the success of the conference. These individuals are: Mary Crovo, Lawrence Feinberg, Ray Fields, Stephen Swearingen, and Daniel Taylor of the NAGB staff; Susan Ahmed, Robert Clemons, Salvatore Corrallo, Arnold Goldstein, Steven Gorman, Jeanne Griffith, Kristen Keough, Andrew Kolstad, Mary Naifeh, Lawrence Ogle, Alex Sedlacek, Sharif Shakrani, Carolyn Warren, Sheida White, and Shi-Chang Wu of NCES staff; Jules Goodison, Eugene Johnson, Steven Lazer, and Paul Williams of ETS; and Luz Bay, Susan Loomis, and Mark Reckase of ACT.

Table of Contents

Page

Foreword i

Acknowledgements ii

Conference Agenda v

Executive Summary ES-1

CONFERENCE PAPERS

Historical Perspective

A Historical Perspective on Setting Standards
Michael J. Zieky 1

Theoretical Perspective

Standards for Reporting the Educational Achievement of Groups
Samuel A. Livingston 39

On the Cognitive Construction of Standard-Setting Judgments:
The Case of Configural Scoring
Richard M. Jaeger 57

Some Technical Aspects of Standard Setting
Huynh Huynh 75

A Conceptual Analysis of Standard Setting in Large-Scale Assessments
Wim J. van der Linden 97

Developing Standards for the National Assessment of Educational Progress

Examinee-Centered vs. Task-Centered Standard Setting
Michael Kane 119

Implications for Standard Setting of the National Academy of
Education Evaluation of the National Assessment of Educational
Progress Achievement Levels
Lorrie A. Shepard 143

Methodological Perspective

Standard Setting--The Next Generation
Ronald A. Berk 161

An Interpretive Approach to Setting and Applying Standards
Pamela A. Moss 185

Page

The Consensus Process in Standards Development
Belinda L. Collins 203

Methodological Issues in Standard Setting for Educational Exams
William A. Mehrens 221

Applications Perspective

Standard Setting From the Perspective of Generalizability Theory
Robert L. Brennan 269

Standards-Based Score Interpretation: Establishing Valid Grounds for Valid Inferences
Samuel Messick 291

Ensuring Fairness in the Setting of Performance Standards
Lloyd Bond 311

Using Performance Standards to Report National and State Assessment Data:
Are the Reports Understandable and How Can They Be Improved?
Ronald K. Hambleton and Sharon Slater 325

Policy Perspective

The Impact of Standards on How Teachers Behave in the Classroom:
Promises and Perils
Phyllis W. Aldrich 347

The Likely Impact of Performance Standards as a Function of Uses:
From Rhetoric to Sanctions
Robert L. Linn 367

Legal Defensibility of Standards: Issues & Policy Perspectives
S. E. Phillips 379

Issues in Standards Setting: Making a Decentralized System Responsive to Centralized
Policy Making
H. Wesley Smith 399

Conference Participants

Program Participants 411

Conference Facilitators 413

Conference Recorders 414

Conference Attendees 415

Hyatt Regency-Capitol Hill
400 New Jersey Avenue NW
Washington, DC

CONFERENCE AGENDA

Wednesday, October 5

9:30 AM	**Registration**
1:00 PM *Plenary Session*	**Welcome** Sharon Robinson, Assistant Secretary, Office of Educational Research and Improvement Mark Musick, Past Chair, National Assessment Governing Board
1:30 - 3:00 *Plenary Session*	**Theoretical Perspectives on Standard Setting for Large-Scale Assessments** Moderator: Joseph Conaty, Office of Educational Research and Improvement Presenters: Samuel Livingston, Educational Testing Service Richard Jaeger, University of North Carolina at Greensboro Huynh Huynh, University of South Carolina Wim van der Linden, University of Twente, Netherlands
3:00 - 3:45	**Dialogue and Open Mike**
3:45	**Break**
4:00-5:30 *Break-Out Groups*	**Discussions on the Theoretical Perspective**
5:30	**Free Time**
6:30	**Dinner Session** Developing Standards for the National Assessment of Educational Progress (NAEP) Moderator: James Popham, IOX/University of California at Los Angeles Emerson Elliott, National Center for Educational Statistics Michael Kane, University of Wisconsin William Randall, Colorado Department of Education Lorrie Shepard, University of Colorado at Boulder

Thursday, October 6

8:00 AM	**Continental Breakfast**
8:30 - 10:00 *Plenary Session*	**Methodological Perspectives on Standard Setting for Large-Scale Assessments** Moderator: Sylvia Johnson, Howard University Presenters: Ronald Berk, Johns Hopkins University Pamela Moss, University of Michigan Belinda Collins, National Institute of Standards and Technology William Mehrens, Michigan State University
10:00 - 10:45	**Dialogue and Open Mike**
10:45	**Break**
11:00-12:30 *Break-Out Groups*	**Discussions on the Methodological Perspective**
12:30 PM	**Lunch**
1:30 - 3:00 *Plenary Session*	**Applications Perspectives on Standard Setting for Large-Scale Assessments** Moderator: Michael Feuer, Board of Testing and Assessment, National Academy of Sciences Presenters: Robert Brennan, University of Iowa Samuel Messick, Educational Testing Service Lloyd Bond, University of North Carolina at Greensboro Ronald K. Hambleton, University of Massachusetts at Amherst
3:00 - 3:45	**Dialogue and Open Mike**
3:45	**Break**
4:00 - 5:30 *Break-Out Groups*	**Discussions on the Applications Perspective**
5:30 PM	**Evening on your own**

Friday, October 7

7:30 AM	**Continental Breakfast**
8:00 - 10:00 *Plenary Session*	**Policy Perspectives on Standard Setting for Large-Scale Assessments** Moderator: Michael Nettles, University of Michigan Presenters: Phyllis Aldrich, WSWHE Board of Cooperative Educational Services Robert Linn, University of Colorado at Boulder Susan Phillips, Michigan State University H. Wesley Smith, Newberg Public Schools
10:00 - 10:45	**Dialogue and Open Mike**
10:45	**Break**
11:00-12:30 *Break-Out Groups*	**Discussions on the Policy Perspective**
12:30 PM	**Luncheon Session**
1:30	**Summary of Break-Out Sessions** James Popham, IOX/University of California at Los Angeles
2:15	**Closing Session, What Have We Learned?** Mark Musick, Past Chair, National Assessment Governing Board Emerson Elliott, Commissioner, National Center for Education Statistics
3:00	Adjournment

Executive Summary

Joint Conference on Standard Setting for Large-Scale Assessments

Linda Crocker / Michael Zieky

INTRODUCTION

Purpose

Our purpose is to provide a summary of the most important information derived from the Joint Conference on Standard Setting for Large-Scale Assessments held in Washington, DC, October 5 - 7, 1994.[1] We do not intend to provide abstracts of the contents of each paper in this summary. Our intent, rather, is to convey the essence of the conference by combining the information into major themes.

Overview

At the conference, papers were presented in each of four "perspectives" on setting standards: theoretical, methodological, applications, and policy. In addition, a paper providing a historical perspective was distributed to attendees. The six major themes we present in this summary cut across the various perspectives. These themes are:

1. Multiple meanings and uses of standards
2. Methods of setting standards
3. New directions and technical issues in setting standards
4. Fairness and validity in setting standards
5. Problems and controversies
6. Areas of agreement in setting and using standards

MULTIPLE MEANINGS AND USES OF STANDARDS

To policymakers, educators, and psychometricians, the term "standards" has multiple and sometimes sharply different meanings. Authors of the papers presented at the conference adopted one or more meanings in framing their remarks. Brennan noted that standards could be considered either as (a) goals declared desirable by an agency or authority or (b) the outcomes of a standard-setting process. A simplistic distinction could be made between defining standards qualitatively or quantitatively. Determining the author's definition of standards is critical to understanding each paper.

Single Definitions

Some authors implicitly or explicitly adopted a single perspective on the meaning of standards and offered their ideas exclusively within that context. But Smith and Aldrich concentrated on how the content of national curriculum standards could affect instructional practice without formally defining their meaning of the word "standards." Collins, on the other hand, cited a definition used by the federal government's Office of Management and Budget which is equally applicable to both educational outcomes and manufactured products, but which was unique among the conference's authors in terms of its breadth.

---

[1] Unless otherwise noted, all references in this summary are to papers contained in Volume II, Proceedings of Joint Conference on Standard Setting for Large-Scale Assessments.

Performance or Content Standards

Other authors reflected on multiple definitions for standards before adopting a single definition for primary use in their papers. Linn identified three types of standards widely used in education: content standards, performance standards, and opportunity-to-learn standards. In this vein, Shepard characterized content standards as curricular frameworks that specify what should be taught at each grade level, while performance standards designate what students must do to demonstrate various levels of proficiency with respect to specific content. Linn, Shepard, and many other authors concentrated primarily on performance standards once their definitions were clearly established. Messick, however, divided his attention almost equally between content standards and performance standards as he discussed validation strategies. Similarly, Phillips, in discussing legal defensibility of standards, seemed to consider both performance and content standards. Finally, Brennan further delineated several types of outcomes of standard setting as performance standards, cut scores on the score scale of an assessment instrument, or exemplars of performance in the form of specific test items or booklets.

Standards as Cut Scores

The perspective adopted by the majority of authors was the view of standards as cut scores (the numeric outcomes of a standard-setting process), benchmarks on a scale, threshold values between contiguous categories, or numeric values that operationalize "how good is good enough" (Livingston, p. 39). The distinction between performance standards and cut scores was elaborated by van der Linden.

> Contrary to a popular point of view, . . . performance standards are verbal descriptions of achievements that form an important step in the process of specification that leads to the domain of test items represented in the test and selects the cut scores. However, once the domain has been realized and the cut scores selected, performance standards lose their operational meaning. From that point on, conversely, the domain of test items and the cut scores define the empirical meaning of performance standards. (p. 98)

This traditional psychometric interpretation of standards as cut scores seemed to undergird the comments of Berk, Bond, Brennan, Hambleton and Slater, Huynh, Jaeger, Kane, Livingston, Mehrens, Shepard, van der Linden, and Zieky.

Multiple Uses

Even when a particular meaning of standards has been established, the intended uses of those standards may still be in question. Messick stressed that different types of standards would have different uses. For example, content standards should be used as "blueprints for teaching and testing" (pp. 294-296) while performance standards, as levels of accomplishment in a specific form, should be "challenges or hurdles" (pp. 296-299). In particular, Linn addressed the multipurpose use of performance standards (exemplification, exhortation, accountability, and certification).

Many authors discussed procedures of standard setting from the perspective of one primary purpose, but the purpose was not necessarily the same across all papers. For example, Phillips, Jaeger, Moss, and Bond all offered examples that represent use of standards for certification of the ability level of individual examinees. However, Livingston reminded readers that sometimes the purpose of standards may be to report information on groups instead of making decisions about individual examinees. Hambleton and Slater, who offered extensive recommendations for standards-based reporting of National Assessment of Educational Progress (NAEP) results, illustrate use of standards in describing group performance. Thus when interpreting an author's comments, readers will find it useful to ascertain (or infer) the author's perspective on both the meaning of standards and the purpose or purposes served by those standards.

METHODS OF STANDARD SETTING

When assessments are used for certification of individuals or for determining the proportions of students judged to be in various classifications, the need for explicit standards is inescapable. A major decision in such an enterprise involves choice of a method for standard setting. Readers interested in acquiring a broad overview of methods of standard setting will especially want to consult the paper by Mehrens. Categorizing and describing various methods of standard setting was a common topic among authors. One widely used scheme was the *examinee-centered* vs. *test/item-centered* method of classification (e.g., Brennan, Huynh, Kane, Livingston, and Mehrens). Both types of standard-setting studies involve the use of expert judges. As noted by Livingston, item-centered methods "involve judgments about the question tasks or problems on the test," while examinee-centered methods "involve judgments about actual test takers" (p. 40). The choice of method dictates substantially different activities, and the differences in results obtained from different methods are nontrivial.

Examinee-Centered Methods

These methods are typically characterized by the borderline-group methods and the contrasting-groups methods. Kane and Zieky are among the authors who described these methods. With the *borderline-group method*, judges identify a group of examinees whose level of achievement is at or near the threshold of minimum acceptable performance. The median (or in some cases the mean) score of this group on the assessment is computed and used as the minimum pass score. With the *contrasting-groups method*, judges identify two groups of examinees, one consisting of persons considered to be masters and the other consisting of persons considered to be nonmasters of the content of interest. The assessment is administered to both groups and the score level that best separates them is determined. Typically, the score that will result in fewest misclassifications is set as the passing score.

Test/Item-Centered Methods

These procedures for standard setting are typically characterized by the Angoff or Nedelsky methods. As summarized by Kane in the Angoff procedure, the judges are asked to imagine an examinee whose ability lies at the threshold of minimally acceptable performance. For each item, the judges then estimate a minimum pass level (MPL) for a group of such minimally qualified examinees on that item (i.e., they estimate what percentage of the group would answer the item correctly). The average MPL over judges is defined as the item MPL, and the sum of the item MPLs is the passing score. The

Nedelsky method is similar but is designed exclusively for multiple-choice items. As described by Mehrens, judges are required to

> look at each item and identify the incorrect options that a minimally competent individual would know were wrong. Then for each judge, the probability of a minimally competent student getting an item correct would be the reciprocal of the remaining number of responses. . . . The expected score on the test for a minimally competent student would be the sum of obtained reciprocals across all items. (p. 227)

These expected scores are then averaged over judges to create the minimum pass score.

Other methods

Although the four procedures described above were those most commonly mentioned by authors, a number of additional standard-setting procedures were mentioned or reviewed in individual papers. For example, Mehrens and Zieky both described methods proposed by Ebel, decision-theoretic methods, and an iterative judgment method proposed by Jaeger. Mehrens used the label *compromise models* to describe standard setting, employing both normative and absolute standards. The three models (Hofstee, Beuk, and DeGruijter) he reviewed were similar in requiring judges to set a passing score (often in conjunction with an appropriate passing rate), but then using that score in conjunction with empirical knowledge of how examinees actually perform on the test. Mehrens also described several recently developed procedures for standard setting that have been used in preliminary work by the NAEP, the National Board for Professional Teaching Standards (NBPT), the National Assessment Governing Board (NAGB), and various state departments of education. Zieky included descriptions of standard-setting procedures used in several different countries.

Comparisons of Standard-Setting Methods

With such an abundance of standard-setting procedures available, a question that logically arises is, "Does it matter which procedure is used?" Quite simply, the answer to this question is, "YES." According to Berk, "Probably the only point of agreement among standard-setting gurus is that there is hardly any agreement between results of any two standard-setting methods, even when applied to the same test under seemingly identical conditions" (p. 162). Likewise, in his "meta-review" of standard-setting research prior to 1992, Mehrens reported that 23 studies had been conducted in which cut scores for the same assessment were established by different procedures. A number of these studies had involved the Angoff method, leading to his report that Angoff cut scores generally are set between Nedelsky and Ebel cut scores, that intrajudge consistency has been reported higher for the Angoff method than for the Nedelsky method, and that the Angoff method typically has lower variance across judges. Findings such as these and the simplicity of implementing the Angoff procedure contributed to the widespread reliance on the Angoff procedure or one of its variations for establishing cut scores in large-scale, high-stakes assessment programs noted by Berk and Kane.

Interestingly, a major impetus for the conference ensued from recent challenges to the widespread acceptance of the Angoff procedure and to item/test-centered standard setting in general. Several writers offered a brief history of this recent debate about the choice of a standard-setting method, and, taken together, their accounts provide a valuable context for understanding this set of papers. Generally, these authors traced efforts by NAGB to establish standards for Basic, Proficient, and Advanced levels of achievement using a version of the Angoff method. Berk noted, and supported with seven citations, that "the test-centered methods used by . . . NAEP to set achievement levels for the 1990 and 1992 assessments were the objects of considerable controversy and, in fact, a heavy barrage of criticism" (p. 162).

The criticisms are well-illustrated in Shepard's paper, in which she described the key findings of the National Academy of Education's (NAE) evaluation of the 1992 NAEP achievement levels. As Shepard recounted, the NAEP originally was charged by Congress to evaluate the effect of extending NAEP to the state level, but

> Because of the salience of standards in educational reform, . . . the National Center for Education Statistics (NCES) asked the NAE Panel to expand its work and conduct an evaluation of the 1992 achievement levels for reading and mathematics. (p.143) The Panel's studies were extensive, including more than a dozen separate field studies and reanalysis of existing data. (p. 145)

Shepard pointed out problems that the NAEP perceived with the pool of NAEP mathematics items and their fit with the standards, with the selection of exemplar items to illustrate the standards, and with the descriptions of the achievement levels.

Comparisons of actual student performance on items to estimated p-values obtained from an Angoff-type procedure led to the conclusion that "judges systematically underestimate the easiness of easy items but overestimate the easiness of hard items" (p. 151). In addition to raising questions about the usefulness of the Angoff procedure, according to Berk, the NAEP recommended that other item-judgment methods should be discontinued in favor of the contrasting-groups approach. These recommendations influenced both the substance and flavor of a number of the papers. For example, Berk and Mehrens mentioned specific recommendations of the NAEP as they reviewed different standard-setting methods. Other authors framed their remarks more generally against the broader canvas of examinee-centered vs. item-centered methods of standard setting for large-scale assessments, but drew specific implications for NAEP assessments as they explored technical issues in standard setting (e.g., Huynh).

NEW DIRECTIONS AND TECHNICAL ISSUES IN STANDARD SETTING

This conference afforded the opportunity for experts in the field to consider new problems in standard setting and proposed solutions as well as issues that might affect use of currently available techniques. The recent surge of interest in performance assessment provided impetus for descriptions of two newly proposed standard-setting techniques. The questions surrounding standard setting for NAEP and the possible use of examinee-centered methods in that assessment spawned exploration of a number of technical and statistical issues. Finally, a growing need for developing recognized criteria or guidelines for the conduct of standard-setting studies was recognized, and evaluative criteria were proposed.

New Procedures

Jaeger and Moss offered descriptions of new procedures for standard setting that are particularly appropriate for performance assessments. Jaeger's empirical test of a policy-capturing study using judgment decision-making (JDM) tasks for judges to evaluate hypothetical profiles of examinee performance on seven performance exercises of the certification assessment for the NBPT illustrates a method that combines aspects of both examinee-centered and item-centered judgments. This paper illustrates application of linear and nonlinear regression models to test whether judges employ compensatory or conjunctive decision-making strategies for setting cut scores on complex performance assessments.

In contrast to Jaeger's psychometric approach, Moss described a standard-setting technique in which judges review a candidate's entire performance across all exercises, determine the candidate's status, and develop a rationale for that determination. In Moss' technique, performance exemplars would serve as standards for different levels of accomplishment in lieu of numeric cut scores. Moss' integrative approach to standard setting was also proposed for application in the context of the certification process of the NBPT standards.

Technical Issues

Various authors noted a sparsity of work in recent years on examinee-centered standard-setting methods and focused their attention on technical issues worthy of attention if such methods were to be used for large-scale assessments. Technical advances in psychometric theory, such as binary and graded Item Response Theory (IRT) models, utility and decision-theoretic models, and generalizability theory were specifically applied to problems in standard setting.

Huynh demonstrated that the minimal examinee sample size needed for standard setting as sample homogeneity and differences between group means vary. Huynh also addressed the ideal item difficulty level for depicting performance of subgroups at specified proficiency levels, and the use of item category boundary estimates or location estimates when calibrations from a graded IRT model are used to create the scale for different proficiency levels. Kane illustrated how the difficulty level of the criterion used to identify contrasting groups affects the threshold score or standard. Kane also discussed the impact, in terms of false positives and false negatives, as the cut score deviates from the mean for the total group, and he described conditions under which false positive or false negative classification errors will increase.

Livingston considered the effect of statistical bias on standards established with item judgment and with borderline-group methods due to effects of regression to the mean and suggested procedures for moderating these effects. The use of continuous utility functions to assign qualifications (proficiency levels) to achievement distributions were proposed by van der Linden. He also described setting "targets" for achievement distributions of groups with respect to these proficiency levels. Brennan applied generalizability theory to conceptualize the problem of how cut scores would vary if the standard-setting process were replicated with different items, judges, and occasions.

"Standards" for Standard Setting

Several authors specifically addressed the growing need for establishing formal criteria, guidelines, or "standards" for practitioners to consider in selection of a method for standard setting and subsequent conduct of the study. For example, van der Linden discussed the following "Standards for Standard Setting" (pp. 107-111): explicitness, efficiency, unbiasedness, consistency, feasibility, and robustness. Berk offered a list of 10 recommendations for standard setting for item-judgment methods, heavily stressing selection and training of judges, as well as a list of "Top 10 Steps to Standard Setting Salvation" (pp. 170-171) which focused on how the judgment process should be structured and carried out to arrive at useful results. Mehrens also indicated that beyond the selection of a particular standard-setting model, other decisions vital to the integrity of the process include

> how the judges should be chosen, how many should be involved, how they should be trained, whether they should meet as a group, how their tasks should be monitored, how separate judge's decisions should be aggregated, what should be documented, and what information should be taken to the policy board that makes the final decision. (p. 247)

He then offered specific guidelines for these practices.

Brennan's list of nine standards for conducting and reporting results of standard-setting studies included appropriate level of aggregation for reporting, estimation of standard errors, intra- and inter-rater reliabilities, reporting of anomalous outcomes, and cautions to users about possible misuses. Messick's criteria for validity of performance standards (structural, generalizability, external and consequential aspects of construct validity) are also applicable to the issue of choice of an appropriate standard-setting method.

FAIRNESS AND VALIDITY IN SETTING STANDARDS

With respect to fairness and validity, authors agreed that it is impossible to set fair standards on unfair assessments, and that it is impossible to set valid standards on invalid assessments. Even if the assessments are fair and valid, however, the standards themselves must also be shown to be valid and fair. Concerns were expressed about the fairness of applying standards when examinees lacked equal opportunities to learn the material being tested. The authors discussed ways to help assure fairness in the setting of standards by following due process constraints and by involving people representing the perspectives of all relevant groups. Several authors indicated that the fairness and validity of standards depend on the consequences of using the standards.

Fairness Issues

The fairness of standards depends first of all on the fairness of assessments. As Bond noted, "it is not possible to set interpretable standards of proficiency on an assessment that is itself 'unfair' or biased against specific groups" (p. 313). Bond also pointed out that proficiency classifications such as "basic" or "advanced" should mean the same thing across subgroups to avoid harmful misinterpretations of assessment results. Livingston acknowledged that issues of fairness in using standards may transcend psychometric issues that apply in setting standards. The results of a standard-setting study provide one source of information to policymakers who may find it "important to avoid a large imbalance between different groups of students in the awarding of some educational benefit. . . . The issue is one of educational or social policy, not of psychometrics" (p. 40).

Opportunity to Learn

Bond used forceful language to warn that "national standards of educational achievement . . . are antithetical to considerations of fairness and equity if there is substantial inequality in educational opportunity among the population of affected students" (p. 319). Bond acknowledged that educational reform cannot be expected to end "poverty, racism, broken homes, and despair" and concluded that

BEST COPY AVAILABLE

a more realistic goal "should be to remove any *official* barriers to educational opportunity for all persons and to encourage universal acceptance of the fundamental premise that all children can learn" (p. 319).

Diversity of Perspectives

Collins stressed that standards should be set "with participants drawn from different key interest groups (so that no single interest dominates the standards development process)" (p. 207). Bond similarly maintained that balance across interest groups in setting standards is important for fairness, and he elaborated: "Diversity of *perspective* should take precedence over ethnic, gender, or 'cultural' diversity, per se, although it is unlikely that the former can be completely assured without some attention to the latter" (p. 316). Messick noted, as did Collins and Bond, that "some means of accommodating diverse viewpoints needs to be considered to make consensus meaningful under conditions of pluralism" (p. 300). Berk stated, "The internal validity of the process hinges on the qualifications of the judges and the procedure used to solicit their judgments" (p. 175).

Process Requirements

Collins discussed ways to achieve fairness in the process of setting standards. She generalized from her experience with standards in industrial settings at the National Institute of Standards and Technology to standards of performance in education. Collins summarized due process requirements for setting standards and isolated five "key principles" for equity: adequate notice of proposed actions; ample provision of opportunities for participation; adequate records of all discussions and decisions; timely distribution of minutes and ballot results; and careful attention to minority opinions (p. 207). Collins cited research showing that standard setting in education generally does not meet the requirements established for standard setting in industry and business.

Legal Issues

Phillips noted that fairness was an important factor in the legal defensibility of standards and stated that fairness in the use of standards requires adequate prior notice to students and school personnel of "the specific knowledge and skills for which students will be held accountable and general guidelines on what constitutes acceptable performance" (p. 382). She warned that the use of new standards will bring renewed attention to differential pass rates. "Although differential performance by itself is not sufficient to invalidate an assessment program, defending such a program against a legal challenge based on alleged discrimination can be costly, time-consuming, and detrimental to public relations" (p. 384). On the same topic, Aldrich cautioned that "fears of litigation about fairness of higher standards may create barriers to the widespread use of high standards" (p. 358).

Dimensionality

Hambleton and Slater wrote that "the validity of the criterion-referenced interpretations depends on the extent to which a unidimensional reporting scale fits the data to which it is applied" (p. 329). Bond also linked standards-based interpretations to unidimensionality: "The setting of a performance standard on a given test implies that a more or less unitary construct is being measured" (p. 316). Messick described concerns that certain assessments may measure multiple dimensions, some of which may be construct irrelevant. "For performance standards to be valid, the increasing achievement levels characterized by such terms as 'basic,' 'proficient,' and 'advanced' . . . should reflect increases in complexity of the construct specified in the content standards and not increasing sources of construct-irrelevant difficulty" (p. 298). A common example of the type of construct-irrelevant difficulty Messick warned against is the use of convoluted language in framing an arithmetic problem.

Validity of Accommodations

Phillips noted a potential conflict between fairness and validity. She warned that accommodations made to increase the fairness of an assessment for people with disabilities may lower the validity of the assessment. For example, a test designed to measure reading comprehension measures a different construct when it is read aloud to examinees. "Drawing the line between a valid and an invalid accommodation requires consideration of the assessment specifications and the uses for the resulting scores" (p. 391).

Construct Validity

Fairness and validity are closely intertwined. Just as Bond began by pointing out that the fairness of the standard is linked to the fairness of the assessment, Messick began by writing that "the validity of these standards cannot be separated from the validity of the assessment itself" (p. 291). Messick elaborated on the requirements of construct validation efforts by raising the issue of the generalizability of standards-based score interpretations across methods of assessment. He indicated that to interpret performance, "in terms of generic constructs requires evidence of generalizability across measurement methods" (p. 296). In addition, he urged that "attention should be paid not just to convergent evidence of consistency across methods, but also to discriminant evidence of the distinctness of constructs within method" (p. 297). Messick warned that the validity of performance standards "is vulnerable to threats of both construct underrepresentation and construct-irrelevant variance" (p. 297). In the former case, nonmasters may meet the standard, and in the latter case, masters may fail to meet the standard.

Validation Evidence

Messick insisted that the construct validity of the assessment and the construct validity of the performance standard "must be evaluated in the same *evidential* terms" (p. 291). Messick stressed that because of the judgmental nature of standards, "their validity depends on the reasonableness of the standard-setting process and of its outcome and consequences, not the least of which are passing rates and classification errors" (p. 300). Kane noted that using performance standards adds validation burdens beyond those required for use of the scores in the absence of standards and wrote, "it is necessary to choose a standard-setting method judiciously and to generate evidential support for the assumptions embedded in the standards-based interpretation" (p. 120). If examinee-centered methods of setting standards are used, Kane argued, "the criterion used in the examinee-centered standard-setting study should be consistent with the test-score interpretation and, therefore, that the test scores should be validated in terms of this criterion" (p. 135). Berk urged consideration of consequential validity in terms of "measures of the political, economic, social, and/or educational outcomes of decisions about examinees" (p. 176). Ultimately, "Decision validity . . . is the acid test of the worth of a standard-setting method " (p. 176). Berk acknowledged, however, the difficulty of obtaining useful criterion data for professional licensing and certification examinations.

Validity and Values

With respect to evaluating the validity of the outcomes and consequences of the use of standards, Zieky maintained "that there is no objective proof of the validity of a standard," because people may "disagree as to whether a particular outcome is appropriate or not" (p. 32). Outcomes considered valid by individuals who hold one set of values may be considered invalid by those with opposing values.

PROBLEMS AND CONTROVERSIES

It became clear at the conference that standard setters continue to disagree about many aspects of their work. No method of setting standards is universally accepted. The Angoff method, which has been the most widely used means of setting standards, was characterized as "fundamentally flawed" by some authors and defended by others. Not all authors agreed that the use of standards would be beneficial, even if the standards had been appropriately set. Authors elaborated on the difficulties of setting standards, noted the legal vulnerabilities of standards, discussed problems in interpreting the results of using standards, and failed to reach consensus on a number of controversial issues.

Lack of Consensus

One of the most pervasive themes of the conference was the constant airing of the unsolved problems and unresolved controversies associated with setting and using standards. As Bond succinctly stated, "The essential problem stems from the simple fact that there is no way to set performance standards that enjoys consensus among measurement specialists" (p. 312). Berk elaborated on the lack of agreement among practitioners: "What optimal combination of ingredients produces an effective standard-setting procedure? . . . We don't know! The problem is that the measurement community has never reached consensus on a set of criteria that can operationally define the 'effectiveness' of any single standard-setting method" (p. 164).

Controversy Concerning the Angoff Method

As noted above, an important source of controversy was Shepard's conclusion that the Angoff method, the most commonly used procedure for setting standards, is "fundamentally flawed" and presents an extremely complex cognitive task that judges are not capable of doing[2] (p. 151). The effect on many standard setters of the characterization of the Angoff method as fundamentally flawed was captured by Berk. "When I first read this . . . it blew me off my beach chair. The Angoff method . . . has been one of the most trusted names in standard setting. And now, it seems as though it's fighting for its life. . . ." (p. 162).

---

[2] Zieky (1994, p. 30) questioned the complexity of the cognitive judgments actually made by Angoff judges. He believes, without proof, that the judges are not actually performing the difficult task of estimating the probability that a member of some hypothetical group of examinees would answer an item correctly. Further, he believes that the judges are, rather, engaged in the much simpler task of expressing their own values concerning how well examinees would have to perform on an item before the judges are willing to say the examinees are competent.

A strongly contrasting view of the Angoff method was offered by Mehrens when he cited a number of studies and concluded,

> The review of the literature suggests the general acceptance of the Angoff method as the preferred model, and this is my recommendation. The recommendation is based on the general reasonableness of the standard set, the ease of use, and the psychometric properties of the standard. (p. 231)

Kane also defended Angoff and other test-centered methods, at least for objective, analytically scored tests. He pointed out that the Angoff method has been used "on a host of licensure and certification tests, as well as on numerous state testing programs, without major complaints from the judges involved" (p. 124). Kane offered rebuttals to a number of the studies upon which Shepard based her conclusion and stated,

> The evidence developed in the five studies of the technical properties of the 1992 NAEP standard setting do not seem to justify the conclusion (Shepard et al., 1993, p. 77) based largely on these studies, "that the Angoff procedure is fundamentally flawed because it depends on cognitive judgments that are virtually impossible to make." (p. 129)

No consensus was formed, but it was not the purpose of this conference to reach closure on the appropriateness of the various approaches to setting standards of performance for the NAEP.

Disagreements on Uses of Standards

The description of disagreements was an ongoing motif at the conference. Aldrich cited authors who disagreed about whether or not it was beneficial to students to set standards at all, regardless of how it is done. Aldrich indicated that those opposed to standards characterize them as a "harmful fantasy" that would take attention away from matters of equity, while those who favor standards see them as a strong motivating force that will help all students. Aldrich herself favored standards as "critical navigational aids," as long as the necessary supporting mechanisms were available (p. 358).

Smith also described controversies about the establishment of standards. He reported that "local people . . . are not prepared to surrender what they believe to be the right of local authority to define the nature of their schools. . . . It is unwelcome news to them that the state will tell them what is to be done . . ." (p. 405). He pointed out that many unresolved issues are likely to lead to future controversies and "setting performance standards will create a great many risks for states, educators, and local policymakers" (p. 406).

Difficulties of Setting Standards

Even if agreement can be reached on whether or not to set standards, there will remain great difficulties in defending any particular standards that are set. Jaeger identified setting standards of performance as examples of "judgment or decision-making tasks (JDM)" (p. 57), and noted that "Responses to JDM tasks, including standard-setting tasks, are . . . responses to problem statements that are replete with uncertainties and less-than-complete information" (p. 58). According to Linn, standard setting looks easy, but is actually hard to do.

> Although both the step of converting a description of a performance standard into a defined range of scores on an assessment and the step of describing what students who meet a given standard of performance actually know and can do may appear straightforward, satisfactory accomplishment of these two steps has proven to be extraordinarily difficult. (pp. 369-370)

The difficulty of setting standards was also clearly articulated by Brennan. "Standard setting is a difficult activity, involving many a priori decisions and many assumptions" (p. 285).

Arbitrary Nature of Standards

According to van der Linden, standards are often perceived as arbitrary.

> The feelings of arbitrariness . . . stem from the fact that although cut scores have an "all or none" character, their exact location can never be defended sufficiently. Examinees with achievements just below a cut score differ only slightly from those with achievements immediately above this score. However, the personal consequences of this small difference may be tremendous, and it should be no surprise that these examinees can be seen as the victims of arbitrariness in the standard-setting procedure. (p. 100)

Legal Vulnerability of Standards

That "feeling of arbitrariness," along with the many unsolved problems and disagreements among recognized experts in the field, intensify many problems in the legal defensibility of standards. Phillips enumerated the conditions "most likely to trigger a legal challenge to a high-stakes, large-scale assessment and its associated standards" (p. 380), including,

> adverse impact on historically disadvantaged groups; use of processes perceived to be unfair, arbitrary or capricious; suggestion that specific attitudes or values are being assessed; failure to provide all accommodations requested by the disabled; and assessing knowledge or skills that examinees have not had the opportunity to learn. (p. 380)

(Note that Glass (1978) forcefully and articulately stated that all standards were arbitrary. Thus, it appears that all standards are vulnerable to legal attack.)

In addition, she warned that performance assessments carry additional vulnerabilities, including lack of experience with the methodology, problems in scaling and equating, lack of generalizability, and the use of fallible human judgments in scoring. High stakes assessments and standards may be attacked under the Fourteenth Amendment to the U. S. Constitution unless adequate prior notice of the new testing requirement has been given, and unless both procedural and substantive "due process" has been followed. That is, both the procedures used to administer the test and the test itself must follow professional standards and be fair[3] to all examinees.

Phillips further cautioned that under First Amendment freedom of speech and free exercise of religion challenges, "parents with definite religious or political views may object to any goal or standard that appears to require the student to espouse a specific belief or point of view" (p. 387). Phillips also alerted standard setters to the need to be very specific about performance requirements and to explicate their assumptions. For example, standard setters "cannot assume that the courts will accept an unstated assumption that the assessment goals intended the measurement to be in English. . . ." (p. 392).

Problems in Reporting Results

Even if standards can survive legal challenge, there remain difficulties in reporting the results. Hambleton and Slater interviewed "policymakers, educators, and people in the media" and found that "many interviewees had problems reading and interpreting the information they were shown" (p. 336) in reports of assessment results in terms of achievement levels. The authors discovered, for example, that only about 10% of the people they interviewed were able to interpret correctly the percentages presented in a table of NAEP results. Even though readers who spent more time on the reports could improve their comprehension, the majority "noted that they did not have the time needed to scrutinize these reports until they could understand them" (p. 340).

Inevitable Nature of Controversy

Are the controversies aired at the conference likely to be resolved over time? Controversial issues have plagued standard-setting activities for years and are not likely to disappear. Zieky, attempting to give a historical perspective, described an "Age of Disillusionment" that began soon after standard setting became a matter of widespread professional attention. He posited that standard setting would necessarily remain controversial because "a standard is *not* a statement of some psychometrically derived truth. A standard is, rather, the embodiment of a set of value judgments. As long as different people hold different values, standard setting will remain a matter of controversy" (p. 29).

AREAS OF AGREEMENT IN SETTING AND USING STANDARDS

Even though controversies and disagreements abounded at the conference, there were some areas of general agreement. Authors agreed that setting standards was a difficult, judgmental task and that the procedures used were likely to disagree with one another. There was clear agreement that the judges employed in the process must be well trained and knowledgeable, represent diverse perspectives, and that their work should be well documented.

---

[3] Because there are many, sometimes contradictory, definitions of "fair" in the context of testing, it is almost always possible to find a published definition under which a test is either unfair to individuals or unfair to members of historically disadvantaged groups.

Standards as Difficult and Judgmental

Mehrens listed several areas of agreement among the experts in standard setting. The first two points of agreement are probably universal among standard setters. "Although the literature on standard setting is inconclusive on many points, there does seem to be agreement that (a) setting defensible standards is difficult," and that "(b) standard setting is a judgmental process" (p. 224). The difficulty of setting standards was commented on by so many authors that it was selected as a component of one of the main themes for this summary (see p. ES-10). Acknowledgment of the judgmental nature of standards is so widespread that Zieky called it "probably the single greatest area of agreement in the history of setting standards" (p. 28).

Lack of True Standards

Authors were in general agreement that standards are constructed rather than discovered and that there are no "true" standards. Mehrens quoted Jaeger (1989, p. 492) that "a right answer does not exist, except perhaps in the minds of those providing judgments" (p. 224). Shepard pointed out that "a number of reviewers" agreed that "true standards do not exist in nature like parameters waiting to be estimated by a reliable statistical procedure" (p. 158). Similarly, van der Linden commented, "some policymakers or educators seem to believe that true standards do exist independently of methods and judges. . . . This view is not correct" (p. 108). Brennan found it to be "particularly important that users be disavowed of any belief that standard-setting outcomes are anyone's 'truth'" (p. 285).

Imperfection of Standard-Setting Procedures

Authors concurred that standard-setting procedures are merely imperfect mechanisms for the collection of information. According to Shepard, "regardless of how statistical they seem, standard-setting procedures are merely formal processes intended to help judges approach their conceptual task in as systematic a fashion as possible" (p. 158). In a similar vein, van der Linden commented on the limitations of standard-setting procedures by noting, "a standard-setting method is nothing but an instrument to elicit responses from subjects from which an estimate of a quantity is inferred" (p. 112). Jaeger pointed out that standard-setting methods "have been ad hoc constructions . . . that are totally devoid of theoretical grounding" (p. 59).

Differences Across Methods

As noted earlier, there was unanimity that the particular method used to set a standard would affect the results. Shepard reported, "the most pervasive finding [of research on setting standards] is that different standard-setting methods produce different results" (p. 156). Mehrens, van der Linden, and Berk, among others, made the same observation. Unfortunately, participants did not agree on what to do about the problems caused by the fact that different standard-setting methods would give different results.

Need for Documentation

Authors believed that all aspects of the standard-setting process should be well documented. Several authors cited relevant entries from the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council in Measurement in Education, 1985) which indicate areas of broad professional agreement. Standard 6.9

specifically requires that when a cut score is used "the method and rationale for setting that cut score, including any technical analyses, should be presented in a manual or report. When cut scores are based primarily on professional judgment, the qualifications of the judges also should be documented" (p. 43). Brennan attempted to encapsulate areas of agreement by suggesting additional standards for standard setting. His first standard also dealt with the need for documentation. "The characteristics of the judges, the items, and the standard-setting process should be described at an appropriate level of detail for the various types of users" (p. 283). Brennan insisted that documentation "should be sufficiently detailed so that an independent investigator could replicate the process" (p. 283). He also insisted that standard setters should document "unanticipated aspects of the standard-setting process, or anomalous outcomes that might affect interpretations" (p. 285). In addition, van der Linden included standards for standard setting in his paper. His first standard, "explicitness," is relevant to the need for documentation as it requires that, "all steps in a standard-setting experiment be based on explicit definitions and procedures" (p. 107).

Possibility of Misunderstanding or Misuse

There was also agreement that, in certain situations, standards could be in error, be misunderstood, be misused, and have potentially harmful consequences. Hambleton and Slater gave examples of difficulties in understanding standards-based score reports. Brennan warned that people who use the results "should be cautioned about any reasonably anticipated misuses of the outcomes of a standard-setting process," and that "the outcomes can be misunderstood" (p. 285). Zieky warned that people who establish a standard should be "ready to modify that standard if experience demonstrates that it is resulting in inappropriate consequences that outweigh the appropriate ones" (p. 32).

Need for Support for Use of Standards

Authors concurred that merely establishing standards of performance will not have the desired effect on educational practices. The development of the essential support structures will be difficult and time-consuming. Linn exemplified the concerns that were expressed:

> For performance standards to have the desired impact on day-to-day classroom activities, they must be internalized by teachers and students. . . . Because the vision of both content standards and performance standards that are being advanced is a radical departure from the vast majority of current daily classroom practices, the transformation cannot be expected to take place overnight. (p. 370)

Aldrich wrote that it would be necessary to change "teacher beliefs and school culture" (p. 358). Smith asserted that it would be necessary to have "a phase-in plan for standards and their assessment. It must address explaining standards-based reform to educators and the public" (p. 404). Smith also indicated that curriculum materials and tools should be available to teachers before standards are assessed.

Judges

Authors agreed that the judges involved in standard setting must represent diverse perspectives, be well trained, and have the required areas of knowledge. Bond expressed as a "cardinal rule" that "those persons who set the standard should be thoroughly knowledgeable of the content domain that is assessed, the population of examinees who take the assessment, and the uses to which the results will be put" (p. 316). Messick stated that informed judgments "require knowledge of the subject-matter domain as well as of the students' levels of developing expertise" (p. 300). Both Mehrens (p. 247) and Berk (p. 169) used the same words to note that judges must be "qualified and credible." Mehrens stated that the judges "must be thoroughly trained to do the job" (p. 248). The agreement among authors about the need for training was made clear by Berk who used colorful language to make the same point: "Train these judges till it hurts" (p. 170). The authors' level of agreement concerning the need for diversity among judges is clearly documented (see p. ES-8) and need not be repeated here.

DISCUSSIONS

After the presentation of each set of four related papers at the conference, discussion sessions were held with the authors. Popham was given the task of orally summarizing the contents of the 16 discussion sessions at the conclusion of the conference. According to Popham, discussants generally agreed that the process of setting standards should be as open and as pluralistic as possible, that it was important to define the purposes of setting standards and to evaluate the consequences in the light of those purposes, that policymakers rather than methodologists should be involved in resolving differences, that equity concerns required involving all concerned constituencies, that more research on examinee-centered methods of setting standards was needed, and that improvements in communicating results to the general public were required.

CONCLUSION

The conference did not result in professional consensus on how standards ought to be set for large-scale assessments. The conference did, however, bring together many of the people most active in the field of standard setting under close-to-ideal conditions to state their views, air their differences, and seek solutions to common problems. Participants certainly gained an understanding of the multifaceted issues involved in setting standards and an awareness of the varied points of view that are held about many of the issues. We hope that this summary will encourage readers to study the papers presented at the conference and that the papers will, in turn, encourage expanded research efforts on the problems that remain in setting standards for large-scale assessments.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Shepard, L., Glaser, R., Linn, R. & Bohrnstedt, G. (1993). *Setting performance standards for student achievement*. National Academy of Education, Stanford, CA: Stanford University.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15*, 237-261.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 485-514). New York: American Council on Education/Macmillan.

A Historical Perspective on Setting Standards

Michael J. Zieky

Executive Director, Educational Testing Service, New Jersey

ABSTRACT

Passing scores were used on tests for centuries with little or no attention paid to how those standards had been set. That long "Age of Innocence" ended with the rapid growth of the criterion-referenced testing movement and the institution of minimum competency/basic skills testing by states and school districts. In response to the needs engendered by those types of tests, there was an outpouring of publications describing systematic methods of setting standards in a brief but intense "Age of Awakening." That age was, however, quickly overlapped by an "Age of Disillusionment" as researchers discovered that the various methods disagreed with one another, and it became abundantly clear that all the methods, no matter how rigorous they might appear, depended at some point on subjective human judgment. The outpouring of methodologies in the Age of Awakening is tapering off, and the shocked outrage of the Age of Disillusionment is gradually cooling as the "Age of Realistic Acceptance" has begun. Even though techniques for setting standards on conventional tests have evolved to the point at which there are several procedures that are in general use, no method of setting standards is universally accepted.

PURPOSE

The purposes of this paper are to describe the history of setting standards with a primary emphasis on the past 40 or so years and to summarize the lessons that have been learned. The focus is on standard setting as the selection of a passing score or a cut score on a test to represent some minimally acceptable level of knowledge, skill, or ability. (Because the various authors cited often use different words to refer to the same concept, the terms "standard," "passing score," "cut score," "cut point," "mastery level," "cutoff score," and the like are used synonymously in the paper.)

This paper will not provide a procedural manual for the application of methods of setting standards, nor will it provide an exhaustive bibliographic review of the voluminous professional literature on the topic. The goal, rather, is to describe salient events in the history of setting standards and to draw lessons from them for use in the future.

OVERVIEW

The paper is organized chronologically, and the history of standard setting has been divided into four stages: the Ages of Innocence, Awakening, Disillusionment, and Realistic Acceptance.[1]

---

[1] Obviously, any attempt to divide the continuous and turbulent flow of events into discrete ages is arbitrary and open to argument. In what is a classic standard-setting problem, the ages have "fuzzy" edges that defy easy categorizations. Nonetheless, the ages have value as organizing criteria, and (as is highly appropriate given the topic of the paper) readers are free to disagree about where the cut points between adjacent ages should have been set.

In the years before 1950, references to methods of setting standards are scarce. Passing scores had been used on tests for centuries, of course, but little or no attention had been paid to how those standards were set. That long Age of Innocence ended with the rapid growth of the criterion-referenced testing movement and the institution of widespread minimum competency and basic skills testing by states and school districts circa 1970-1980.

In response to the needs engendered by those events, starting in the early 1970s, there was an outpouring of publications describing systematic methods of setting standards in an intense Age of Awakening. That age was, however, very quickly overlapped by the Age of Disillusionment (circa 1978) as researchers discovered that the various methods disagreed with one another, and it became abundantly clear that all the methods, no matter how rigorous they might appear, depended at some point on subjective human judgment. Ironically, it was only after measurement theorists and educators began applying systematic methods of setting standards and trying to set standards in rational ways that strong criticisms of standards emerged.

The outpouring of methodologies in the Age of Awakening is tapering off with a mild flurry of activity concerned with newer modes of testing. The warning cries of the Age of Disillusionment are gradually weakening as the Age of Realistic Acceptance has begun. We have come to realize that the researchers who warned that there is no really objective way to set standards are correct, but we have also come to realize that there is nothing wrong with using judgments appropriately when there is a need to set standards. Unfortunately, we still have great difficulty in agreeing on exactly what procedures would be appropriate in certain standard-setting efforts.

Even though techniques for setting standards on conventional tests have evolved to the point at which there are several procedures in general use, no method of setting standards is universally accepted, and no specific procedures for implementing any method are universally accepted. We are beginning to understand the implications of the fact that a standard is not a statement of some objective truth. A standard is based on the values of a group of people, and, as long as different people hold different values, standard setting will remain controversial.

THE AGE OF INNOCENCE

Standard setting goes back a long time. Documentation of probably the first quantitative standard-setting session for what would now be called an authentic, city-wide, performance-based, high-stakes test can be found in a reference available in most hotel rooms in the United States.

> And the men turned their faces from thence, and went toward Sodom: but Abraham remained standing before the Lord. And Abraham drew near, and said, Wilt thou also destroy the righteous with the wicked? Peradventure there will be fifty righteous within the city; wilt thou also destroy and not spare the place for the fifty righteous that are therein? . . . Shall not the Judge of all the earth do right? And the Lord said, if I find in Sodom fifty righteous within the city, then I will spare all the place for their sake. And Abraham answered and said . . . Peradventure there shall lack five of the fifty righteous: wilt thou destroy all the city for the lack of five? And he said, If I find there forty and five, I will not destroy it. And he spoke unto him yet again, and said, Peradventure there shall be forty found there. And he said I will not do it for forty's sake. And he said unto him . . . Peradventure there shall be thirty found there. And he said, I will not do it, if I find thirty there. And he said . . . Peradventure there shall be

> twenty found there. And he said, I will not destroy it for twenty's sake. And he said, Oh let not the Lord be angry, and I will speak yet but this once: Peradventure ten shall be found there. And he said I will not destroy it for ten's sake. (Genesis 18:22-18:32)[2]

Why did Abraham decide to stop his requests for concessions at the presence of 10 righteous people in the city rather than 11 or 9 or 5 or 15? The passage illustrates a method of setting standards that is strikingly similar to what later came to be known as "counting backwards." The desired performance level is assumed to be high, "but it is quickly recognized that perfection is impossible and concessions must be made for mental infirmity, clerical errors, misinformation, inattention, and the like. Just how great a concession has to be made becomes distressingly arbitrary" (Glass, 1978, p. 244).

In one aspect, standard setting has not advanced since the time of Abraham. Diverse panels of subject matter experts interacting with highly trained consultants armed with powerful personal computers may be used instead of one man standing before the Lord armed with a sense of justice, but all methods of setting standards remain "arbitrary" in the sense that judgment is involved at some stage of the process.

Although the origins of large-scale, sophisticated testing can be traced back to China thousands of years ago, relatively unstructured oral examinations predominated in the United States until about the middle of the 19th century (U.S. Congress, Office of Technology Assessment, 1992). Standards of a sort existed in the form of passing scores on these tests, but the standards were highly dependent on the judgment, whim, biases, and state of mind of the examiner at the time the test was administered. Standards cannot be consistently defined or applied in the absence of tests that are standardized so that all examinees are faced with comparable tasks under comparable conditions.

Later in the 19th and early 20th centuries, formal written testing began to be increasingly used both for instructional management and for accountability purposes as the country became more urbanized and schools had to cope with increases in enrollment and increases in diversity of the student population.

Ability tests were among the earliest examples of tests that were standardized for reasonably large populations. The 1908 Binet-Simon intelligence scale appears to have used the counting backwards method of setting standards in determining mental levels. In the instrument, the various tasks to be performed were grouped into clusters identified by the ages at which children were expected to be able to perform them. "A child's mental level was determined by the highest age at which he or she passed all (or all but one)" (Thorndike & Lohman, 1990, p. 13).

---

[2] William Angoff preferred to quote the following passage as an early example of standard setting.

> And Jephthah gathered together all the men of Gilead, and fought with Ephraim. . . . And the Gileadites took the passages of the Jordan before the Ephraimites: and it was so, that when those Ephraimites that were escaped said, Let me go over; that the men of Gilead said unto him, Art thou an Ephraimite? If he said, Nay; Then they said unto him, Say now Shibboleth: and he said Shibboleth: for he could not frame to pronounce it right. Then they took him and slew him. (Judges 12:4-12:6)

Angoff said that the passage from Judges was preferable because it contained a real "cut" score (personal communication, circa 1973). (All biblical quotations are from *The Holy Scriptures*, New York: Hebrew Publishing Co., 1939.)

Ebel (1965) stated that "In the early years of the century almost all marking was in percents. . . . A definite percent of 'perfection' usually between 60 and 75 percent was ordinarily regarded as the minimum passing score" (p. 406). How were those standards selected? As Zieky (1987) noted, "The most common methods of setting standards are those based on tradition ('The passing score is 70 percent because the passing score has always been 70 percent.') or on power ('The passing score is 80 percent because I say so.')" (p. 2). Another example of the way standards were set is provided by an accusation concerning the influence of quotas on the passing scores established for tests used for entrance to secondary schools in England (circa 1920): "All that was involved was the drawing of a line at the requisite point, decided by the number of secondary school places actually available, and declaring that children below that line had failed to qualify" (Sutherland, 1984, p. 188).

The large-scale use of objective questions and standardized tests was begun by the armed forces to classify recruits during World War I. The new testing technology was adopted for use in education with great expectations for its effectiveness, in part because it would allow the use of standards of accomplishment. In 1924, Ellwood Cubberly, Dean of the School of Education at Stanford University, wrote an introduction to a text that explained the intricacies of the new mode of testing. Note the distinction made between what would now be called norm-referenced and criterion-referenced interpretations of scores, and the delight in the fact that the new measures would allow the development of standards.

> Up to very recently our chief method for determining the efficiency of a school system was the method of personal opinion. . . . Relatively recently the method of comparison was introduced. . . . It is evident that this is a much better method than the one of personal opinion. Its chief defect, though, lies in that the school system studied is continually compared with the average or median of its size and class. In other words, the school system is continually measured as against mediocrity, when as a matter of fact the average or median school system may not represent a good school system at all. . . .
>
> Still more recently, and wholly within the past decade, a still better method for the evaluation of the work which teachers and schools are doing has been evolved. This new method consists in the setting up, through the medium of a series of carefully devised "Standardized Tests," of standard measurements and units of accomplishments for the determination of the kind and the amount of work which a school or school system is doing. . . .
>
> To the teacher it cannot help but eventually mean not only concise and definite statements as to what she is expected to do in the different subjects in the course of study, but the reduction of instruction to those items which can be proved to be of importance in preparation for intelligent living and future usefulness in life. . . . To the citizen the movement means the erection of standards of accomplishment which are definite, and by means of which he can judge for himself as to the efficiency of the schools he helps to support. For the superintendent it means . . . the establishment of standards of work by which he may defend what he is doing. (Monroe, DeVoss, & Kelly, 1924, pp. xi-xiii)

How were the standards to be set? The authors offered no specific methodologies, but did give general advice that standards must be reasonable and efficient. "To be reasonable a standard must be such that it can be obtained by pupils, under school conditions, and with an appropriate time expenditure" (Monroe et al., 1924, p. 419). Thus there was to be an explicit normative aspect to standards. An efficient standard "must represent a degree of ability which equips pupils for meeting present and future demands with a high degree of efficiency" (p. 420). The reference to "present and future demands" indicates that there was also to be an absolute aspect to standards.

Buck (1977) provided an example of the role of the federal government in setting standards in the 1940s.

> The Veterans Preference Act of 1944, requiring that veterans be given points to be added to earned passing marks, mandates the use of cutting scores for Federal examinations. Rule II, Section 4, 1947 of the Civil Service Rules states that Civil Service ratings will have a maximum rating of 100 . . . and a minimum eligibility rating of 70. (p. 4)

Buck noted that the cut scores were not necessarily set at 70 percent, however. The cut scores were set where desired, and raw scores were scaled in a way that made the cut point equal a scaled score of 70. How were the actual cut scores set? Buck's paper included a description of an essentially normative process based on how many job openings were expected, how many people were likely to apply, and what the score distribution was likely to be.

Gulliksen (1950) discussed the problem of setting standards in his work, *Theory of Mental Tests*. His focus was on transforming a raw score scale to some standard scale with "critical limiting values." He gave as examples of such values the grades used in certain colleges on a "scale from 100 (representing a perfect score) to 65 or 70 (representing the failure line)," and Civil Service ratings in which "70 is defined by regulations as the mark to be assigned to the lowest acceptable performance" (p. 265). His advice on selecting the critical level is certainly reasonable, but not very specific.

> In determining this point all relevant factors must be considered, such as the probable difficulty level of the examination, the standards it is necessary to maintain, and the number and percent of candidates above or below this critical point. (p. 266)

Gulliksen (1950) also warned against the practice of looking for a gap in the score distribution and setting the passing score just above the gap because "such gaps are purely accidental and should be ignored in favor of more rational considerations in determining the critical points" (p. 266).

Gulliksen (1950) did give very specific and useful advice on setting a standard on a test when the test was used to predict a criterion with a known critical level such as obtaining passing grades in college. He outlined an essentially decision-theoretic approach, though he did not use that label.

> A cutting score just below F [a point on a graph] would mean that the lowest persons accepted would have a 50-50 chance of being above the critical criterion level. . . . The decision to move the cutting score away from point F depends on judging either that the need for additional persons is sufficiently urgent to justify accepting those who are more likely to fail than to qualify, or that we can afford to reject a group that has a

> better than even chance of success in order to reduce the total number of failures. (p. 295)

Gulliksen (1950) went on to demonstrate that the approach to setting a cutting score could be quantified if it were possible to determine the cost of selecting a person who would fail and the gain from selecting a person who would pass, and he listed the equations needed to perform the calculations. The idea of weighting the losses of each type of misclassification has been carried forward into currently used, test-based methods of setting standards. See, for example, the discussion of the contrasting-groups method in Livingston and Zieky (1982).

Just a few years later in 1954, Nedelsky's article, "Absolute Grading Standards for Objective Tests," was published in *Educational and Psychological Measurement*. Nedelsky's work on standard setting has been called "the oldest procedure that still enjoys widespread use" (Jaeger, 1989, p. 495). The method is worth examining not only for its own sake, but also because it is a precursor of other test-based methods of setting standards currently in use. Nedelsky described in detail a systematic method for setting standards on multiple-choice tests. He introduced the concept of the "F-D student" on the borderline between failing and barely passing. Such students should be able to reject the options in multiple-choice items attractive only to F students. They will, however, have to select at random from among the remaining responses. The reciprocal of the number of remaining responses to an item is the probability that F-D students will respond correctly to the item. The most probable mean score of the F-D students is the sum of the reciprocals. Nedelsky's article contained very clear, step-by-step directions for implementing the standard-setting model.

Nedelsky (1954) seemed to anticipate later critics of his method by explicitly acknowledging that reality did not correspond well to his model. "It should be clear that 'F-D students' is a statistical abstraction. The student who can reject the F-responses for every item of a test and yet will choose at random among the rest of the responses probably does not exist" (p. 5).

He did, however, report the results of a pilot study demonstrating that the method using the abstraction of the F-D student worked reasonably well. Though it is not possible to show direct influence, concepts described by Nedelsky are found in more recent descriptions of methods of setting standards. The concept of the F-D student on the borderline between passing and failing corresponds to Angoff's (1971) "minimally acceptable person," to Ebel's (1972) "minimally qualified (barely passing) applicant," and to the members of the "borderline group" described by Zieky and Livingston (1977).

One good method to trace the way a topic is viewed by the members of a profession over time is to see how that topic is treated in major texts at different periods. A very informative view of the transitions in the history of setting standards can be found in the treatment of the issue in the three editions of *Educational Measurement* published in 1951, 1971, and 1989, edited by E. F. Lindquist, R. L. Thorndike, and R. L. Linn, respectively. Investigation of how the topic is treated over time in such a widely used, comprehensive text gives a clear view of the accepted canon of knowledge concerning standard setting at the time when each edition was compiled. By chance, the dates of the three editions happened to fall at times that afford clear views of how the treatment of standard setting changed over time.

Flanagan's chapter, "Units, Scores, and Norms" in the first edition of *Educational Measurement* (1951) contained a clear description of what later came to be known as criterion-referenced measurement. Even though he never used the phrase "criterion-referenced," Flanagan carefully distinguished between information derived from test scores in terms of test content and information derived from test scores in terms of rank in a specified group. Of importance to the history of setting standards is the fact that he explicitly associated content-based score interpretations with setting standards of achievement.

> The most basic type of information obtained from tests refers to the individual's knowledge and ability with respect to the content itself. This information tells us directly what the individual did with respect to the questions and problems set by the test. It contrasts with the other types of information in which the individual's score is described by comparison with other scores obtained on the same test. . . . Examples of descriptive information in terms of content include such statements as, "This individual knows all of the single digit pairs of addition combinations," . . . or "This individual can translate a typical section of 100 lines from a French newspaper with fewer than five errors." . . . The statement with respect to the ability to translate materials from a French newspaper provides a practical report regarding performance at a specified level of difficulty. Such statements have direct meaning without the scores of other individuals. This type of information lends itself especially well to setting standards of achievement. (p. 700)

Flanagan (1951) had earlier distinguished between norms and standards in a way that sounds quite familiar to those aware of the debate about whether large-scale assessments such as the National Assessment of Educational Progress (NAEP) should focus on what students are actually able to do or on what judges believe the students should be able to do. (See, for example, U.S. General Accounting Office, 1993a.)

> Norms describe the actual performance of specified groups of individuals. "Standards," on the other hand, are *desirable*, or desired, levels of attainment. . . . For example, the norm of quality of handwriting at the end of fifth grade is 50 on the Ayres Handwriting Scale. On the other hand, studies indicate that a satisfactory standard of legibility on this same scale is approximately 70. (Flanagan, 1951, p. 698)

Flanagan (1951) showed a clear awareness of the problems often associated with setting standards of performance. His description of the reactions of standard setters to an unexpectedly low passing rate is also likely to sound very familiar to people involved in setting standards today.

> In terms of content a definite score has been obtained which may be related to standards established by the examiner who prepared the questions. It may be regarded by this examiner as a good or poor score on the basis of his judgments of the difficulties of the items and the expected performance of those taking the examination.
>
> In many cases such judgments are difficult to make and are not related to the realities of the situation. Frequently, when teachers, civil service examining boards, or similar groups discover that all of those examined have fallen below the standard they originally set for the examination, they revise their judgments regarding the content of the test materials and either apply a correction factor to the scores or modify the standards in terms of the obtained scores. (p. 701)

In spite of the clear descriptions of content referencing, the careful distinction between norms and standards, and the acknowledgment that the judgments involved in setting standards are difficult and often unrealistic, Flanagan offered no help in how to go about making the necessary judgments.

The first edition of Ebel's book, *Essentials of Educational Measurement*, entitled *Measuring Educational Achievement* (1965), gave no specific advice on how to set passing scores. Indeed, Ebel discouraged the practice.

> Educational achievement is a continuous variable. No great gulf is fixed between those who pass and those who fail until the final decision is made. Any cutting score is at least partly arbitrary, and some failures, as well as some passes, can properly be attributed to chance, to bad or good luck, as the case may be. There is no way of avoiding the arbitrariness or the influence of chance in the ultimate decision, but there is no point in multiplying the problem by making pass or fail decisions on each test. (p. 80)

THE AGE OF AWAKENING

The second edition of *Educational Measurement* appeared 20 years after the first edition, in 1971. It happened to fall at the dawn of an age of explosive growth in standard-setting methodologies. In the interval between the two editions, criterion-referenced measurement had emerged under that label. The first usage of the term "criterion-referenced" is usually attributed to an article by Glaser in 1963, though Glass (1978) referenced an earlier mention in a chapter by Glaser and Klaus (1962). However, many of the concepts involved, such as the distinction between standards and norms and the focus on test content, had been articulated earlier (see, for example, Flanagan, 1951).

The growth of the criterion-referenced testing movement had a tremendous impact on the history of setting standards. Even though Glaser did not use the word "criterion" to refer to a cut score, cut scores were used in the operational implementation of almost all criterion-referenced tests in schools (Hambleton & Rogers, 1991). Criterion-referenced tests became so ubiquitous, they were perceived as a threat by some proponents of norm-referenced tests. By 1977, for example, the publishers of a major norm-referenced test battery felt constrained to combine a defense of norm-referenced interpretations of scores with a mild attack on absolute standards.

> Many of the proponents of criterion-referenced tests seem to find it necessary to try to discredit norm-referenced tests. We regret this attempt to belittle the value of norms. Scores on tests like ITBS or ITED can be interpreted in either the criterion-referenced or the norm-referenced sense. We have tended to emphasize norm-referenced interpretations, however, because the derivation of reasonable expectations--independent of norms--is generally very difficult. (Peterson, 1983, p. 211)

In one of the great ironies of the history of setting standards, the so-called Angoff method, which was destined to become an extremely well-known and widely used method of setting standards, was described in a mere footnote to Angoff's chapter, "Scales, Norms, and Equivalent Scores," in the second edition of *Educational Measurement* (Thorndike, 1971).[3] The footnote explained a "slight variation" on the process Angoff had very briefly described in the text as a "systematic procedure for deciding on the minimum raw scores for passing and honors" (p. 514). In a section of the text devoted to descriptions of various score scales, Angoff very concisely described a method for setting standards.

> Keeping the hypothetical "minimally acceptable person" in mind, one could go through the test item by item and decide whether such a person could answer correctly each item under consideration. If a score of one is given for each item answered correctly by the hypothetical person and a score of zero is given for each item answered incorrectly by that person, the sum of the item scores will equal the raw score earned by the "minimally acceptable person." (p. 514)

In the footnote, Angoff (1971) expanded on the process somewhat to allow probabilities rather than only binary estimates of success or failure on each item.

> A slight variation of this procedure is to ask each judge to state the *probability* that the "minimally acceptable person" would answer each item correctly. In effect, the judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities, or proportions, would then represent the minimally acceptable score. (p. 515)[4]

The second edition of *Educational Measurement* gave evidence of the increased interest of the measurement profession in setting standards in addition to the introduction of what came to be known as the Angoff method. The text included a section entitled "Methods of Setting Cutoff Scores" in Hills' chapter, "Use of Measurement in Selection and Placement" (Hills, 1971).

In the context of admissions, Hills (1971) described a method of setting a cutoff score by "predicting categories from measurements" that is quite similar to what later became more widely known as the contrasting-groups method. It involves dividing examinees into two groups on the basis of some criterion of success or failure external to the test.

> One procedure is to draw the overlapping frequency distributions of the predictor variable for the successful and unsuccessful groups.
>
> If these distribution curves intersect, the point of intersection (at which point an individual has a 50-50 chance of success) might be chosen as the cutoff point. (p. 699)

---

[3] As an additional irony, Angoff insisted that the method should properly be attributed to Ledyard Tucker (Angoff, personal communication, circa 1982). Also documented in Jaeger, 1989, p. 493.

[4] In the light of the controversies that have erupted over whether the word "could," "should," or "would" should be used in describing the performance of the minimally competent individual when applying the Angoff method, it is very interesting to note that Angoff used the word "could" in the text and "would" in the footnote. When Livingston and I were writing *Passing Scores*, I asked Angoff which word was correct and should be used in the application of the method. He replied that he did not think it mattered very much (personal communication, 1982).

In addition, Hills (1971) discussed decision-theoretic approaches to setting standards and noted the difficulty of applying the methods because the "payoff function" is often not known.

In the context of placement, Hills (1971) discussed quota-based methods of setting cut scores that are of little interest. However, he also described methods of setting cut points in which data on predicted criteria were available.

> These, then, are two more principles for setting cutoff points, i.e., (a) setting the cutoff according to a predetermined probability of obtaining a specified grade and (b) setting the cutoff at a predetermined grade level and assigning the student to the highest-level course for which his predicted grade reaches that level. (p. 712)

Also in the context of placement, Hills (1971) discussed decision-theoretic approaches to setting standards and showed how to select the cut score on a placement test when aptitude-treatment interactions were found.

The same edition of *Educational Measurement* contained a chapter by Glaser and Nitko (1971), "Measurement in Learning and Instruction," in which cut scores are addressed as well. The authors described methods for treating the establishment of a cut score on a test as a hypothesis testing exercise in which the probability of a correct mastery classification can be calculated given, among other data, the minimum acceptable proportion of tasks mastered in the domain. The problem remains, however, of determining the domain cut score in the first place.

In any case, by the time of the publication of the 1971 edition of *Educational Measurement*, it was clear that the setting of standards had become a matter of concern to members of the measurement community, and researchers were groping for appropriate methodologies. It is interesting to note that neither Angoff nor Hills nor Glaser and Nitko referenced Nedelsky's article, "Absolute Grading Standards for Objective Tests." The article was clearly relevant to the problems addressed in their chapters and had been printed in a major journal about 17 years earlier (Nedelsky, 1954).

It was between the publication of the second edition of *Educational Measurement* in 1971 and the publication of the third edition in 1989 that the knowledge explosion of what may be termed the "Age of Awakening" in standard setting took place.

As late as the 1977 edition of their introductory text, *Measurement and Evaluation in Psychology and Education*, Thorndike and Hagen could ask, "Where does the criterion level representing satisfactory mastery come from?" and answer simply, "The criterion level of mastery seems often to be set quite arbitrarily on the basis of someone's 'professional judgment' " (p. 116). In 1988 a similar question was raised in another introductory measurement text, *Psychological Testing* (Cohen, Montague, Nathanson, & Swerdlik, 1988), but the answer given was quite different. "How should cutoff scores in mastery testing be determined? How many test items . . . are needed to demonstrate mastery in a given field? The answer to these and related questions could be the subject of a text in itself" (p. 85). Clearly perceptions of the amount of information available about the field of standard setting had changed in the interval between the two publications.

In Ebel's text, *Essentials of Educational Measurement* (1972), the author described a method of setting standards that later entered the canon of widely cited procedures and was named after him. He was among the authors who explicitly acknowledged that "Determination of a minimum acceptable performance always involves some rather arbitrary and not wholly satisfactory decisions" (p. 492). In his chapter, "Standard Scores, Norms, and the Passing Score," Ebel described five methods of setting standards, referring to them as "the first approach, the second approach, etc." The author described as the first approach what is essentially the counting backwards method for use on tests of minimal essentials.

> Theoretically, the passing score on such a test should be a perfect score. In practice, of course, to insist on a perfect score as the minimum passing score would be almost to guarantee that no one would pass. Minimum essentials are not obvious enough to all, the item writer's ability to test them unequivocally is not perfect enough itself, and the examinee's performance is not flawless enough to make a perfect score a minimum score. There needs to be some margin for errors on all counts. (p. 492)

Ebel (1972) next explained as a second approach a method of setting standards that has not achieved widespread popularity. It involved determining the "ideal mean score" on a test, which is a "point midway between the maximum possible score and the expected chance score." The passing score should then be set "midway between the ideal mean score and the expected chance score" (p. 492).[5] Ebel acknowledged the inappropriateness of the method for tests in which the obtained mean and minimum scores departed from the ideals. He also pointed out a weakness of the method.

> The items may be more difficult, or less difficult or less discriminating, than the test constructor intended. Whether an examinee passes or fails a specific test may be determined by the questions in the test rather than by his level of professional competence. (p. 493)

In another great irony in the history of setting standards, what came to be famous as Ebel's method was proposed by the author as merely a way of overcoming a weakness of the second approach. Ebel (1972) did not even count it as a separate method of setting standards in its own right. "The . . . weakness of this approach can be overcome to some degree by the derivation of the passing percentage from a subjective analysis of the relevance and difficulty of each item in the test" (p. 493).

Ebel (1972) included a table illustrating how to divide the items into four categories of relevance, "essential, important, acceptable, questionable" and into three categories of difficulty, "easy, medium, and hard." He demonstrated how to calculate the standard based on the number of items in each cell of the matrix formed by the crossing of the relevance and difficulty categories by estimating the expected success rates on those items of a "minimally qualified (barely passing) applicant" (p. 494).

His third approach was merely normative and involved picking a percentage of examinees to fail. The fourth approach was much more promising and foreshadowed later so-called compromise methods of setting standards. (See, for example, Beuk, 1984, and De Gruijter, 1985.) Ebel (1972) proposed combining the first and second approaches to obtain the advantages of using both normative and absolute methods. He gave as an illustrative example, "the passing score might be defined as 75 percent correct responses, provided at least 60 percent of the examinees but not more than 80 percent of them exceed this score" (p. 495). The fifth approach used the performance of certified practitioners on the test as a basis for selecting a minimum score to certify new practitioners.

---

[5] At first glance, this may appear to be a method of setting standards that does not require judgment. Further reflection, however, shows that deciding to set the cut "midway" is indeed a matter of judgment. Why not two-thirds of the way or 13/18 of the way? It is also clearly a matter of judgment to have selected the "ideal mean" as one of the reference points. Why not 1.96 standard errors of measurement above the ideal mean or one standard deviation below it?

Hambleton and Novick (1972) demonstrated the use of decision theory and Bayesian estimation to help in determining a student's mastery status. The use of decision theory requires the assignment of values to the losses associated with false positive and false negative errors of classification, or at least the specification of the ratio of the values of the losses. It also requires that some "arbitrary threshold score used to divide individuals into the two categories" be provided (pp. 4-5). The Bayesian estimation allows the use of "collateral information contained in test data of other students" or possibly "the student's other subscale scores and previous history" to help determine the student's mastery status (p. 6). In a statement that is as true now, more than 20 years later, as it was then, Hambleton and Novick wrote, "At this stage of the development of a theory of criterion-referenced measurement, the establishment of cut-off scores is primarily a value judgment" (p. 4).

By 1973, there were many requests by school district personnel for practical help in setting standards on basic skills tests. In response to those requests, the Educational Testing Service (ETS) ran a series of workshops on the topic, and Zieky (1973) compiled descriptions of methods of setting standards under seven labels. The first two, "Traditional" and "Judgment Uncontaminated by Data" were described as in common use, but were not recommended. "Judgment Based on Inspection of Items" described the Angoff method. "Judgment Based on Inspection of Required Sequential Behaviors" was a variant of the Angoff method for use in instructional hierarchies. In this method, judges determine whether or not success on each item is required for the work in the next level of the hierarchy. "Empirical Use of Normative Data" was essentially the borderline-group method and involved administering the test to a group of students judged to be minimally competent. "Empirical Use of a Predicted Variable" was merely a way of transferring a known cut point on some criterion measure to the predictor test. "Empirical Use of Preselected Groups" was quite similar to what was later known as the contrasting-groups method. It involved using information independent of the test to select masters and nonmasters and then finding the test score that best separated the two groups.

Another early entry in the information explosion surrounding standard setting was Millman's 1973 article, "Passing Scores and Test Lengths for Domain-Referenced Measures." He provided one of the early summaries of known methods of setting standards, including those based on quotas, the performance of people already certified, and evaluations of test items. Millman described Ebel's method (in press at the time) and quoted relevant aspects of the Angoff and Nedelsky methods. He gave some sensible advice based on decision-theoretic principles concerning adjusting the cut score up or down depending on the costs associated with remediation compared to the costs of moving a student too quickly through the material to be learned.

Of greater interest to the history of setting standards was the fact that Millman also presented a very useful table showing the percentage of students expected to be misclassified on tests of different lengths with different passing scores, depending on the students' true level of functioning. The table showed, for example, that a passing score of four out of five items would misclassify about half of the students whose true level of functioning was 70 percent correct. The quantification of misclassification rates helped make clear that the imposition of an artificial dichotomy on an essentially continuous distribution of ability would result in errors of classification.

By 1976, there was enough interest in standard setting to warrant an article in the *Review of Educational Research*. Meskauskas (1976) divided standard-setting methods into two broad categories.

The author described models in which mastery was assumed to exist along a continuum and "state" models in which mastery was treated as "all-or-none." The continuum models included familiar methods such as those of Angoff and Ebel, along with the much less familiar Kriewall binomial-based model. Kriewall's method required "assuming a standard of performance and then evaluating the classification errors resulting from its use. If the error rate is inappropriate, the decision maker adjusts the standard a bit and tries his equations again" (Meskauskas, 1976, p. 139). The state models described by Meskauskas (Emrick's mastery testing evaluation model and Roudabush's dichotomous true score models) have not found widespread use, probably because few people believe that mastery and nonmastery exist as true dichotomies.

By the mid 1970s, the interest in standard setting had grown among members of the measurement community to the point that Glass (1978) complained, "at an AERA symposium entitled Criterion-Referenced Testing, four of the five papers were essentially psychometric treatments of the cut-off score problem" (p. 242). At the concurrent meeting of the National Council on Measurement in Education, there was a symposium devoted to the topic of measurement issues related to performance standards. Shepard's 1976 paper, "Setting Standards and Living with Them," is of interest in the history of setting standards because it focused on practical advice for using standards rather than on methodological advice for how to set standards. As did many authors, Shepard acknowledged the subjective nature of standards. She also pointed out that absolute standards are influenced by norms. The judgments involved in setting standards are tempered by the judges' experience, "But what do we mean by experience except imperfect norms? . . . There are some advantages to judges being influenced by formal and explicit norms rather than unsystematic and internal ones" (p. 6). She suggested that standard setting should be iterative and based on the effects of using the standard. She further noted that judges ought to be provided with normative information to help them make reasonable decisions and that all relevant audiences ought to be involved in the process.

Federal regulations concerning employee selection set requirements for cut scores. The references to "normal expectations" and "adverse impact" indicate a belief that there would be normative influences on standards.

> Where cutoff scores are used, they should normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the work force. If other factors are used in determining cutoff scores, such as the relationship between the number of vacancies and the number of applicants, the degree of adverse impact should be considered. (U.S. Department of Justice, 1976, p. 51753)

A method of setting standards based on the "familiar known groups validation procedure" was proposed by Berk (1976). The concept of setting standards based on the score distributions of two groups recurs several times in the history of standard setting. See Gulliksen (1950), Hills (1971), Livingston (1976), and Zieky and Livingston (1977) for examples of variations on this theme. Berk's method involved administering the test to both instructed and noninstructed groups and finding the point on the score scale that best separated the two. He pointed out that there would be two types of errors of classification and said that the optimal cut score "maximizes the probability of correct decisions . . . and minimizes the probability of incorrect decisions" (p. 6).

Livingston (1976) proposed using methods of "stochastic approximation" to establish a cut score. The relevant situation is one in which a standard is to be set on a relatively inexpensive written test, based on achievement on a more expensive performance measure on which success or failure can easily be judged. In this example of the use of two groups to set a standard, Livingston used performance on the expensive measure to establish group membership. He described the general technique:

> 1. Select any person. Record his written test score and measure his actual performance.
>
> 2. If the first person succeeds on the performance measure . . . choose next a person with a somewhat lower written test score. If the first person fails on the performance measure, choose next a person with a somewhat higher written test score. (pp. 2-3)

By continuing the process, scores on the expensive criterion measure are concentrated in the portion of the range where they are needed to set the cutoff score. The score that best separates the two groups can be determined. One of the methods was described for a more general audience in Livingston and Zieky (1982).

The history of standard setting was affected by the efforts of researchers who began performing studies of the various methods that had been proposed. In one of the early studies, Andrew and Hecht (1976) compared the application of Ebel's method and Nedelsky's method of setting standards and found that the results of the two methods were widely discrepant. They concluded that "the specific techniques employed in setting an examination standard may be a more powerful determinant of the standard than any other variable" (p. 49). This was disconcerting news because it cast doubt on the meaningfulness of standards. If it is assumed that there is some correct standard and if two methods designed to find that standard disagree, then one method, or both, must be wrong. Andrew and Hecht's findings were corroborated a number of times for different methods of setting standards. (See Jaeger, 1989, pp. 497-500, for a description of the outcomes of many such studies.)

Even though the history of setting standards had advanced to the point at which some reasonably rational methods for setting standards were available, many standard setters ignored the methods in practice. Shimberg (1982) reported the results of a study done in 1976 of the pass/fail rates on licensing tests for 12 skilled trades in four states. For every occupation, the study found that pass rates were high when employment was high and that pass rates were low when employment levels were low. As Shimberg noted, the "study points up possible weaknesses in the way passing grades are established" (p. 54).

The history of standard setting was strongly influenced by the extremely rapid growth of district and statewide minimal competency testing. "For example, from 1972 through 1985 the number of state testing programs grew from one to 34" (Madaus & Kellaghan, 1991, p. 14). People who were not necessarily measurement specialists were faced with the task of setting standards on mandated high-stakes tests. They had no choice but to comply with the laws mandating testing, and many needed help with the practical logistics of setting defensible operational standards.

In 1977 Zieky and Livingston described methods of setting standards in a *Manual for Setting Standards on the Basic Skills Assessment Tests*. An important aspect of the manual was that it was written for school district personnel who had to face the job of setting standards rather than for graduate students or measurement professionals. Detailed step-by-step instructions and sample forms were provided. The manual described how to apply Nedelsky's method, Angoff's method, what the authors called the borderline-group method, and the contrasting-groups method. The borderline-group method depends on judges to identify minimally competent examinees who then take the test. Their median score is the estimate of the standard. The contrasting-groups method requires judges to identify examinees who are likely to be masters of the objectives measured by the test and examinees who are likely to be nonmasters of those objectives. After the test is administered to the two groups, the score level that best separates them can be determined. The numbers of misclassifications of each type that occur can be estimated, and the standard can be adjusted to minimize the type of misclassification that is judged to be worse.

Practical advice was offered on topics such as obtaining a diverse group of judges and the amount of time that should be set aside for the process of setting the standard. Zieky and Livingston (1977) were explicit about the judgmental nature of setting standards. They also stressed that the effects of the use of the standards should influence the standard-setting process.

> It is extremely important to realize that all methods of setting standards depend on subjective judgment in some point in their application. There is no good way of setting standards just by plugging numbers into a formula.
>
> It is also important to realize that standards are never set in a vacuum. The placement of the standard will affect how the school district allocates its resources. It will affect the professional lives of teachers and administrators, and it will certainly have an impact on the lives of the students taking the tests. Awareness of the potential effects of the standards-setting process should influence the judgments that enter into that process. (p. 2)

Buck (1977) made a useful contribution to the history of standard setting by discussing the cut score issues related to test bias and adverse impact in the context of "some practices and procedures in the Federal Government." He pointed out that ". . . in Federal employment examinations cutting scores are required for both legal and administrative reasons" (p. 1). Speaking from a perspective other than the one shared by many measurement professionals, Buck stated, "the use of cutting scores can help to reduce administrative problems, public relations problems and the overall cost of the testing and decision making processes" (p. 2). He provided a very informative example of the factors that influenced standard setting in a program that received 200,000 applications for about 12,000 job openings.

> The first year of the program the cutting score was set realistically in terms of the number of jobs to be filled. The office was subsequently inundated with letters from Congressional representatives on behalf of their constituents, college professors on behalf of their students, and from the applicants themselves. It was necessary to set a lower cutting score. (p. 4)

Jaeger (1978) reviewed a number of methods of setting standards (adopting a fixed percent correct, using the performance of some external normative group, stochastic approximation, decision theoretic, Nedelsky's method, Ebel's method, and Angoff's method) and found each one either flawed or inapplicable to the problem of setting standards on a state-wide high school competency test. He therefore proposed a new method based on the use of multiple groups of stakeholders and iterative procedures that came to be known as Jaeger's method.

For the high school competency test, Jaeger suggested using very large samples of registered voters, high school teachers, and high school administrative and guidance staff for a total of 700 people. He proposed dividing them into separate, homogeneous groups of 50, having them take the test and

BEST COPY AVAILABLE

providing keys to the items. The judges would then be asked to answer one of two questions: "Should every high school graduate be able to answer this item correctly?" or "If a student does not answer this item correctly, should (s)he be denied a high school diploma?" Jaeger noted that neither question depends on the concept of minimal competence. Then each judge's cut score is determined, and data showing the group's cut scores are displayed along with item data. The process of judging is reiterated. Normative data are shared showing projected passing rates, and the judgments are reiterated. The median judgment for each of the three original groups is calculated and the standard is set equal to the minimum of the medians.[6]

THE AGE OF DISILLUSIONMENT

As noted previously, the Age of Awakening continued as the Age of Disillusionment began, and they have run concurrently ever since. Some authors appear to be living in the Age of Awakening and others in the Age of Disillusionment.

One important landmark in the history of standard setting was the devotion of the entire Winter 1978 issue of the *Journal of Educational Measurement* to the topic. A special journal issue was a clear signal that standard setting had become a topic of great importance to members of the measurement community. It is of great interest to note that much of the issue (six out of eight articles) was devoted to a discussion of whether or not setting standards is a justifiable activity, rather than to descriptions of methods of setting standards or to summaries of research findings related to standards.

Glass (1978) forcefully and articulately stated that all standards were "arbitrary."

> I have read the writing of those who claim the ability to make the determination [of] mastery or competence in statistical or psychological ways. They can't. At least, they cannot determine "criterion levels" or standards other than arbitrarily. The consequences of the arbitrary decisions are so varied that it is necessary either to reduce the arbitrariness, and hence the unpredictability, or to abandon the search for criterion levels altogether in favor of ways of using test data that are less arbitrary and, hence, safer. (p. 237)

Glass (1978) went on to classify methods of setting standards into six categories (Performance of Others, Counting Backwards from 100%, Bootstrapping on Other Criterion Scores, Judging Minimal Competence, Decision-Theoretic Approaches, and Operations Research Methods) and to demonstrate why he believed that the methods in each category were arbitrary and, therefore, fatally flawed. Glass unambiguously stated his conclusion: "Setting performance standards on tests and exercises by known methods is a waste of time or worse" (p. 259).

A companion article by Burton (1978) corroborated the main points made by Glass.

> I presented three possible procedures for setting performance standards beyond the classroom. Theoretical principles were rejected because learning hierarchies have never been established and other theories at present seem too limited. Expert judgments were

---

[6] Selecting the minimum of the medians implies that a judgment was made that it was worse, in this context, to fail a person who deserved to pass than it was to pass a person who deserved to fail.

> rejected beyond the classroom level because informed professional decisions require a level of information not available beyond the individual classroom. Finally, "minimal competency" techniques were rejected because real-life successes have many potential causes. No single skill is so essential that it can be defined as necessary for survival. (pp. 270-271)

Rebuttals were given in the same issue of the *Journal of Educational Measurement* in a set of articles by Scriven (1978), Hambleton (1978), Block (1978), and Popham (1978). Scriven decried the lack of research on setting standards and wrote:

> To put the matter bluntly, the answer to Glass does not lie in present practice (nor does it lie in surrender) but rather in elements missing from the whole picture, including better procedures for calibrating and training judges, for synthesizing subtest scores, and especially in needs assessment. (p. 274)

Scriven's call for more research is still appropriate today. In effect, Glass has not yet been answered. There continue to be more questions than answers with respect to the variables that affect the outcome of a standard-setting study.

Popham and Hambleton agreed with Glass that standard-setting methods were indeed "arbitrary," but they pointed out that arbitrary can mean "judgmental" as well as "capricious."

> The cornerstone of Glass's attack on the setting of standards is his assertion that such standards are set *arbitrarily*. He uses that term in its most pejorative sense, that is, equating it with mindless and capricious action. But while it can be conceded that performance standards must be set *judgmentally*, it is patently incorrect to equate human judgment with arbitrariness in this negative sense. (Popham, 1978, p. 298)

Continuing professional interest in setting standards was demonstrated by the National Council on Measurement in Education's publication in 1979 of *Practices and Problems in Competency-Based Measurement*. The book contained a chapter on standards that included papers by Jaeger (1979), Shepard (1979), and Conaway (1979). The overlapping of the Age of Awakening with the Age of Disillusionment is strikingly clear in these documents. As more information was gained about the processes of setting standards, more was learned about the problems inherent in those processes.

Jaeger (1979) pointed out 13 possible threats to the validity of standards. He described seven categories of standard-setting methods and indicated which threats to validity affected each type of method. A study of Jaeger's list of threats to validity is a useful, but depressing activity for anyone contemplating the task of setting standards.

1 = Bias in setting domain standard due to inadequate domain definition

2 = Random error among judges who set domain standard

3 = Inappropriateness of item sampling procedure, bias error in sample standard

4 = Inadequate item sample size, random error in sample standard

5 = Judgment bias in the consideration of individual tasks

6 = Lack of representativeness in the criterion group used to determine percentiles

7 = Inadequately large criterion group, resulting in random error

8 = Lack of representativeness in the examinee group, leading to bias error

9 = Inadequately large examinee group, leading to random error

10 = Model bias due to inaccuracy in loss judgments, false positives or false negatives overvalued

11 = Bias error due to invalidity of domain identification

12 = Error due to inconsistency of domain-criterion relationship

13 = Bias error due to invalidity of model for domain-criterion relationship. (p. 57)

In the light of Jaeger's daunting list, it is not surprising that one of Shepard's (1979) recommendations for setting standards was to "avoid setting standards whenever possible." She offered suggestions designed to make standards appropriate, but warned, "these precautions cannot compensate for the errors implicit anytime a continuum of performance or skill development is represented as a black and white dichotomy" (p. 67).

In keeping with the other papers, Conaway (1979) stated:

> A review of the literature shows that a definitive set of practical procedures for setting standards in various types of competency-based programs does not exist. . . . Although increased attention has been given to the issue in the past few years, it is apparent that practitioners cannot obtain validated standard-setting procedures, either off-the-shelf or out of the literature. (p. 72)

The continued strength of the criterion-referenced testing movement was illustrated by the publication of a special issue of *Applied Psychological Measurement* in 1980 devoted to the topic. As might be expected, standard setting was the focus of several articles in the issue. Shepard (1980) reviewed the state of the art in setting standards. She divided the known methods of setting cut points on a continuum of ability into five categories and discussed their characteristics. In a more novel section of the paper, Shepard discussed the selection of standard-setting methods for specific uses such as diagnosis, certification, and program evaluation. She suggested not using standards for program evaluation because "they obscure performance information about individuals along the full performance continuum" (p. 464). For pupil diagnosis, Shepard encouraged teachers to keep in mind both absolute and normative conceptualizations of mastery. For pupil certification, Shepard made the same suggestion in far stronger terms.

> At a minimum, standard-setting procedures should include a balancing of absolute judgments and direct attention to passing rates. All of the embarrassments of faulty

> standards that have ever been cited are attributable to ignoring one or the other of these two sources of information. (p. 463)

In the same issue, van der Linden (1980) provided a highly quantitative article that Livingston (1980) called "a thorough and precise mathematical presentation of decision theory as applied to educational testing" (p. 577). Livingston went on to note that the rigor of the article may discourage some test users from the use of decision theory. He continued, though, that decision theory "is really common sense expressed in mathematical language, and using contrasting groups data to set a cutoff score is one of the simpler applications of decision theory" (p. 577).

A distaste for standards continued among some members of the profession. Green (1981) wrote "A Primer of Testing" for readers of *American Psychologist* who were "unfamiliar with the essentials of psychological tests" (p. 1001). Concerning standards, Green suggested avoidance to the novices.

> The fact that one item does little by itself is the main reason professional testers eschew fixed cutting scores. Sometimes, as in competency tests, the situation forces establishment of a cutscore, but since one item may well make the difference between passing or failing the cutoff, fixed cutting scores are to be avoided whenever possible. (p. 1005)

Another demonstration of the strong link between criterion-referenced testing and standard setting was provided by *Criterion Referenced Measurement: The State of the Art*, edited by Berk (1980). Hambleton's (1980) chapter entitled "Test Score Validity and Standard-Setting Methods" makes an important contribution by reinforcing the connection between standards and validity. There is a listing of what had become the usual methods and advice on how they should be implemented. There is also a consideration of the types of decisions to be made on the basis of the test scores to help determine the extent and sophistication of the standard-setting process that will be required. "Analysis of the decision-making context involves judging the importance of the decisions that are to be made using the test, the probable consequences of those decisions, and the costs of errors" (p. 108).

By 1981, the knowledge explosion with respect to standard setting in the Age of Awakening had reached the point at which Hambleton and Powell were able to say:

> Certainly there is no shortage of (1) discussions of issues associated with standard setting . . . (2) methods for standard setting . . . and (3) reviews of those methods . . . . What does seem to be in short supply are guidelines to help groups work their way through the issues and technical matters which must be addressed in selecting and implementing a standard setting method and finally setting a standard. (pp. 1-2)

The authors provide those guidelines in the form of extensive lists of questions that standard setters should consider about important and practical issues such as available resources, applicable laws, selection of judges, use of norms, analysis of data, and so forth. The shift to practical issues of implementation rather than a further compendium of methods indicated that, at this stage in the history of standards, issues of how to carry out methods were coming to the fore.

Such practical issues were the focus of *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests* (Livingston & Zieky, 1982). It was written to offer advice to people who were not measurement professionals but were faced with the task of setting standards.

The manual made the usual warnings about the judgmental nature of standards and the need to verify absolute standards with normative information. The authors gave detailed instructions for using the Angoff, Ebel, Nedelsky, borderline-group, contrasting-groups, and up-and-down methods of setting standards. They also provided advice for choosing among the methods and stated a very strong preference for the contrasting-groups method. Their statement concerning the contrasting-groups method has taken on additional relevance as performance-based measures have become more widely used.

> If you can have the judges actually observe the test-takers' performance or samples of their work, we recommend using the contrasting-groups method. This situation will occur fairly often with essay tests, hands-on performance tests, etc. . . . The contrasting-groups method has the strongest theoretical rationale of any of the methods we have presented: that of statistical decision theory. It is the only standard-setting method that enables you to estimate the frequencies of the two types of decision errors. (p. 53)

The manual also warned would-be standard setters that "Choosing the passing score on a test often leads to controversy" (p. 55). It also discussed questions that should be considered prior to setting standards and offered a number of practical "helpful hints."

By the early to mid 1980s, methods of setting standards had become an accepted aspect of what people trained in measurement were expected to know. Indeed, by that stage in the history of setting standards, knowledge of the topic was no longer considered esoteric. School teachers as well as psychometricians were learning how to set standards. In what Nitko (1983) described as a "basic textbook for an introductory educational testing and measurement course for elementary and secondary school teachers" (p. v), descriptions were given of normative methods, the counting-backwards method (Nitko descriptively called it the "feet on the desk" procedure), the contrasting-groups method, and Nedelsky's and Ebel's methods of setting standards. (Angoff's method was mentioned only in a footnote, which somehow seems appropriate given its origin.) In addition, Nitko provided a copy of Hambleton's (1978) practical guidelines for the determination of cutoff scores.

In a higher level text intended for "students of test theory," Crocker and Algina (1986) devoted an entire chapter to setting standards. They described what had become by then the ordinary methods for setting standards. In their review of the empirical research, they noted that different standards are set using different methods or different judges. They made a very important point concerning the limited role of psychometrics in setting standards. "It is imperative to recognize that although standard setting is an important psychometric problem, it is not solely a technical issue. The consequences of appropriate or inappropriate standards for individuals, for institutions, and perhaps for society as a whole must be considered" (p. 419).

Normative information can help determine whether or not standards are sensible. If, for example, there is reason to believe that most of a cohort of examinees are competent, a standard that fails most of the examinees would be considered unreasonable. An important contribution to the history of setting standards was made by Beuk (1984) and De Gruijter (1985) who suggested systematic methods for reaching compromises between absolute and relative (normative) standards. Considering both types of information in setting a standard can help avoid the establishment of unreasonably high or low standards.

By the time the third edition of *Educational Measurement* was published in 1989, information about methods of setting standards had become so widespread that Jaeger was compelled to provide a taxonomy of standard-setting methods in his chapter, "Certification of Student Competence," to help keep order among the methods. He cited at least 18 different methods with "additional variations" on those methods. He gave detailed explanations of the Angoff, modified Angoff, Ebel, Jaeger, Nedelsky, borderline-group, and contrasting-groups procedures. He described research findings related to the methods. He tabulated the results of about a dozen studies showing that the different methods of setting standards disagreed with each other.

As a very rough indicator of the change in magnitude of the amount of information on standard setting made available to readers, consider that in the 1971 edition of *Educational Measurement* about 10 lines (in a footnote) were devoted to what became known as Angoff's method. In the 1989 edition, over 110 lines were devoted to Angoff's method in Jaeger's chapter, in addition to the approximately 600 lines of text devoted to other methods and issues related to setting standards.

The three editions of *Educational Measurement* were used as convenient markers for stages in the history of setting standards, but, of course, work on setting standards did not stop with the publication of the third edition. An entire issue of *Educational Measurement: Issues and Practice* was devoted to the subject in 1991. With respect to the history of setting standards, it is very informative to compare the topics of the articles with the focus of earlier work in the field. Jaeger (1991) wrote about the problems of selecting appropriate judges. He discussed issues of what kinds of judges to obtain and how many judges were needed. Mills, Melican, and Ahluwalia (1991) discussed how best to explain the standard-setting process to judges and how to help the judges reach a definition of minimal competence. Reid (1991) also commented on how to train the judges who participate in the studies. Plake, Melican, and Mills (1991) described factors that influence judges when they set standards. Geisinger (1991) wrote about the issue of adjusting the value derived in the standard-setting study to reach an operational cutoff score. Rather than describing methodologies or arguing about whether standards should be set, these authors explored the nuances of how best to carry out the process. At this point in the history of setting standards, there was general agreement on the essentials of the process. Even students in introductory courses were exposed to them. The measurement professionals had turned their attention to an exploration of the details and toward the challenges posed by the shift to what some were calling the "newer modes" of assessment.[7]

RECENT CHALLENGES

Even though many of the problems associated with setting standards on traditional tests have not yet been solved, we are beginning to address the additional complications of applying standards to constructed-response tests, performance tests, and computerized adaptive tests (CATs).

---

[7] One interesting feature of doing the background reading for this paper was to see material written in the first quarter of this century refer to objective tests as the "new" modes of assessment, and material written in the last quarter of the century refer to nonobjective tests as the "new" modes of assessment. I came away with the feeling that progress in certain aspects of educational measurement might best be described by Ecclesiastes (1:9): "The thing that hath been, it is that which shall be; and that which is done is that which shall be done: and there is no new thing under the sun."

Faggen (1994) described four methods of setting standards that were appropriate for constructed-response tests: (1) the benchmark method, (2) the item-level pass/fail method, (3) the item-level passing-score method, and (4) the test-level pass/fail method.

In the benchmark method, judges study the scoring guides and "benchmark papers" selected to illustrate the performance expected at various levels of the score scale. The judges select papers at the lowest level they consider acceptable. The process is iterative, and data are shared and discussed. The average score of the lowest acceptable papers is the recommended passing score.

In the item-level pass/fail method, judges read a number of papers without knowledge of the grades assigned. They judge each paper as passing or failing. Results are collated and discussed. Judges have the opportunity to change their ratings. The data then are used to estimate the probability that a paper at each score level is acceptable. If the two errors of classification are considered equally harmful, the point at which the probability is 0.5 is the recommended standard. This is similar to the contrasting-groups method.

In the item-level passing-score method, judges study the scoring rules and descriptions of performance at each score level. They then estimate the average score that would be obtained by a group of minimally competent examinees. The average estimate across judges is the recommended standard. This is similar to Angoff's method.

As might be expected, the test-level pass/fail method is similar to the item-level pass/fail method except that the judgments are made at the level of the entire set of examinee responses to the constructed-response questions in a test. (For a one-item test, of course, the two methods are identical.) Faggen described an interesting adaptation of the method she called the "see-saw" in which the score of the response that the judge is asked to rate next depends on the judgment made for the previous response. This allows a rapid targeting of papers near the judge's estimate of the pass/fail point.

One problem in setting standards on measures of complex, multidimensional behaviors is that it is impossible to capture all the aspects of an examinee's performance in a single score. One example of such a measure is the Praxis III performance test designed to help provide information useful in the evaluation of entry-level teachers. The measure provides evaluations on 19 separate criteria. Setting a single overall cut point on the total score using one of the traditional standard-setting methods may not be satisfactory because certain judges may be concerned about the "profile" or distribution of scores across criteria used to reach the overall score. As Jaeger (1994) wrote:

> All of the most prominent performance-standard-setting procedures (e.g., Angoff, 1971; Ebel, 1972; Jaeger, 1982; Nedelsky, 1954) have in common the expectation that the tests to which they will be applied are unidimensional, and, correspondingly that the items that compose the tests contribute to a summative scale. (p. 3)

The judges may find some profiles of scores unacceptable even if the mean scores averaged across the criteria are above the overall passing point. To illustrate the problem, consider two profiles of four scores each. The judges may not wish to consider a profile of 1.0, 1.0, 3.5, 3.5 to be equivalent to a profile of 2.0, 2.0, 2.5, 2.5, even if the means are identical. The judges may feel that a minimally qualified teacher should not receive any scores of 1.0 (the lowest possible score), or no more than one such score. They may feel that a score of 1.0 is acceptable on some criteria but not on others. They may feel that a score of 1.0 is acceptable on some criteria only if it is balanced by a score of 3.5 on certain other criteria. How can this complexity be captured in the establishment of a standard?

One approach suggested by Jaeger (1994) with respect to a collection of performance exercises developed for the National Board for Professional Teaching Standards was the use of "judgmental policy capturing." The process is complicated and iterative, but, in essence, judges are shown many profiles of scores and asked to give an overall rating to each profile based on their professional judgment. Then, mathematical models are used to "capture" the implicit policies that influenced the judges' decisions. Indices of model fit can be used to see how well the process worked.

In addition to devising methodologies for constructed-response tests and multidimensional performance tests, standard setters have faced the challenge of devising appropriate methodologies for setting standards on computerized adaptive tests. A CAT form is constructed from a pool of questions as each person takes the test. The computer selects questions that cover the specified content at an appropriate level of difficulty for each individual. There is no particular problem in setting standards on a CAT if it is possible to use one of the methods based on judgments of examinees such as the contrasting-groups method or the borderline-group method. The results of those methods can be applied directly to scores derived from administrations of the CAT.

The methods of setting standards based on judgments of test items cannot be applied directly to a CAT, however. The quantity estimated by methods of setting standards based on judgments of items is the average-number-right score of a group of people on the borderline between competence and incompetence. Those number-right scores cannot be applied directly to a CAT because if a CAT is working perfectly, *all* test takers will answer about half of the items correctly regardless of differences in ability.

How can standards be set on a CAT if it is not practical to obtain judgments of the competence levels of test takers? Matha Stocking (personal communication, 1992) proposed using judgments of test questions by employing the concept of an "exemplar" test form. An exemplar form is a traditional linear test whose items have been calibrated to the same metric as the CAT item pool. The exemplar test is built to the same content specifications as the CAT, using the same types of questions. One reasonable exemplar would be the set of questions administered on the CAT to a test taker of average ability. Once the exemplar has been constructed, standards can be set on the exemplar using any of the generally accepted standard-setting procedures designed for traditional tests. The application of the method would result in a standard expressed as a number-right score on the exemplar test. Once a number-correct score has been established on the exemplar test, it can be mathematically transformed into an estimated ability on the underlying Item-Response-Theory metric. An examinee's adaptive test score can then be determined to be below or above this passing point.

A novel method that takes advantage of the nature of adaptive tests was proposed by Walter Way (personal communication, September 1994). He suggested having judges take the adaptive test and respond to items as they believed a minimally competent examinee would respond to the items. That is, each judge would decide whether or not a minimally competent (borderline) examinee would be able to respond correctly to an item. If the item appeared to be so easy that a borderline examinee would have a high probability of responding correctly to the item, the judge would respond correctly to the item. If the item appeared to be at a difficulty level that would cause the minimally competent examinee to have a low probability of answering correctly, the judge would respond incorrectly to the item. An interesting aspect of the method is that the item selection algorithm would tend increasingly to select items that were near middle difficulty for the minimally competent examinee being simulated by the judge. That allows more information to be gathered in the region of the cut score than is likely to be obtained with a traditional linear test. At the conclusion of the CAT, the resulting ability-level score should be an estimate of the ability level of a minimally competent examinee as determined by the judge. Averaging ability levels across judges would result in an estimate of an appropriate standard.

STANDARD SETTING IN OTHER COUNTRIES

A search of the professional literature on testing and standards in other countries returned a great deal of information about national testing programs, but very few specific facts about how standards were set. I tried to get detailed information by communicating directly with colleagues in other countries, and they were quite helpful and informative. Based on their replies to my request for information, and further based on the review of the literature, I have a strong sense that there are no magic standard-setting methodologies waiting to be discovered overseas. Practices differ in various countries, as might be expected. Some are quite elaborate and impressive. Others are less extensive. In many ways, their experiences with standards are similar to ours.

I did, however, receive the impression that the use of standards provokes much less controversy in some other countries than it does in the United States. One reason is that the cut scores are simply not released in some countries. Another reason may be that the tests are subjectively scored to a greater extent than is the case here, and experienced examiners can adjust grading standards to keep passing rates fairly constant from year to year. This consistency in passing rates was corroborated by Eckstein and Noah (1993).

> Success rates in examinations differ quite markedly across nations, but each nation appears to regard its own particular rates as somehow part of the natural order of things. For example, in France from year to year the pass rate on the *baccalaureate* has moved only a few percentage points above or below two-thirds; and the general belief in France is that it could hardly be otherwise. (p. 188)

Broadfoot (1992) supports their view of the revered status of the *baccalaureate* in France. The passing rates are likely to be held constant. "Given the very strong support for the *baccalaureate* as an anonymous, external examination . . . and as a guarantee of equality against the influence of teachers' values and schools of different status, all attempts at reforming the '*bac*' have so far largely failed" (p. 323).

Normative data were also used to maintain passing rates in England and Wales.

> It is not widely recognized outside the education service that although the grading of individual attainment in CSE and GCE examinations is based upon the quality of work presented, grade boundaries are strongly influenced by statistical norms.
>
> For example, grades A to C in GCE O-level are awarded to a fairly constant proportion of candidates every year. (Department of Education and Science, Welsh Office, 1982, p. 10)

Reddaway (1988) reported on how the University of Cambridge Local Examinations Syndicate set standards on the GCE Advanced-Level Technology examination. His report supports the belief that normative data affect the standards that are set.

> Standards fixing is based substantially upon the professional judgment of the senior examiners who will be required . . . to suggest cutoff points for the pass/fail line and the minimum mark for the award of the highest grade. Where an examination and the nature of the entry to it have changed little since the previous year, the percentages reaching the recommended marks will be compared with the results of previous years as supplementary information. . . . Admittedly much is undefined and much rests upon the experience of the senior examiners and the continuity that this provides, but it enables the system to cope with the setting of complex tasks to the candidates and to react to the need to change the curriculum . . . while still attempting to maintain constant standards. (p. 29)

Van Krieken (1987) indicated that normative data influence cut points for the central examinations for secondary general education in the Netherlands.

> When enough answer sheets have been scored or when the scores of an adequate sample of teacher-marked essay papers have been received, the intended cutting-off point is reconsidered. When a paper has proved to be exceptionally difficult or easy, the intended cutting-off point will be adjusted. . . . In this way a compromise has been reached between norm- and criterion-referenced testing. (p. 37)

Eckstein and Noah (1993) went on to state that the observed passing rates in various countries "are the result of choices, some deliberate, some less conscious, made within educational, examination, and political institutions" (p. 188). Indeed, the authors referred to passing rates as "artifacts of national policy."

In China, university entrance examination scores are "either above or below the cutoff point, adjusted each year to produce the number of freshmen desired by the . . . authorities" (Eckstein & Noah, 1993, p. 172). In Japan, scores are reported on the entrance examination, but no indication is given of passing or failing. (That, of course, is the system used for college entrance examinations in the United States as well.)

In Alberta, Canada, all grade 12 students are required to take at least one examination to graduate. Examinations are offered in a number of subjects. The examinations are developed by the Department of Education with a great deal of review by classroom teachers. Cut points are set at fixed percentages correct, in what appears to be an arbitrary selection of 80% correct for "excellence" and 50% correct for passing. However, the effects of the fixed cut points are mitigated by adjustments based on judgment.

> Teachers are . . . involved in confirming cutpoints . . . and in grading open-ended questions. . . . Teachers involved in grading determine whether or not the content and difficulty of questions accurately reflect curriculum standards. Scores are compared to previous examinations, and adjustments may be made to control for increases in the level of difficulty. (U.S. General Accounting Office, 1993b, p. 20)

Gunter Trost, Director of the Institute for Test Development and Talent Research, responded to my request for information on how standards were set in Germany on large-scale examinations. Even though there is a common "minimal requirement" on the *Abitur,* it appears to mean different things in different jurisdictions.

> We do have written examinations at the end of secondary school (as part of our "Abitur" exam), but they are non-standardized and differ from State to State, in most of the States even from school to school. In order to graduate from upper secondary school, the students must achieve an average mark no lower than 4.4 (on a scale of 1.0--best--through 6.0--totally insufficient). The average mark is calculated from marks in various written exams . . . some oral exams and achievement in the last two years of school. This "minimal requirement" is agreed upon by all States (and accepted by the public), but even the way of computing the average mark (e.g., by allocating different weights to the respective marks) may differ from State to State. (personal communication, August 1994)

In Ireland, the Leaving Certificate Examination (LCE) is offered to pupils who have reached the age of 16 and have followed an approved course in school. "The results of the LCE are among the principal means employed by Irish society for admitting persons to university, teacher training, the civil service, and many other careers" (Greaney & Kellaghan, 1979, p. 132). The examinations, controlled by the Department of Education, generally consist of constructed-response items and extended essays, and are centrally scored by teachers who are carefully trained and supervised. (They do not score the papers of their own students.) Supervisors check the distribution of each examiner's grades to maintain equivalence. Grades are awarded on a 13-point scale associated with percent-correct scores.

How are cut scores set on the Leaving Certificate Exam? Thomas Kellaghan, Director of the Educational Research Center, Saint Patrick's College, responded to my query: "I don't know if anyone could answer precisely your question about setting cut scores on public examinations. No details are published" (personal communication, August 1994).

That the cut scores are adjusted on a more or less normative basis can be inferred from Kellaghan's description of the centralized scoring system.

> When all scores are in, final adjustments may be made to grade boundaries. These will take into account the distribution of grades in a subject in preceding years and the distribution of grades in other subjects in the current year. Unless good reasons are apparent to the Chief Examiner, he/she will attempt to maintain uniformity across subjects and across years. (personal communication, August 1994)

He closed by writing that the examination system was not the subject of a great deal of criticism. "Most people (teachers, students, parents, [university officials], politicians) regard them as a fair and impartial means of assigning educational benefits. People particularly like the idea that the system is free of any kind of personal influence or corruption" (personal communication, August 1994).

Ingemar Wedman, Professor of Education and Educational Measurement, University of Umea, answered my request for information about standards in Sweden with some very interesting information.

> In late Spring this year the Parliament decided to move from a norm-referenced grading system to a criterion-referenced grading system without paying any attention to the problem of how to set the different standards. . . . Most of the work of setting standards is left . . . to the individual teacher to do. It's today difficult to predict how this work will be carried out. (personal communication, August 1994)

Graham Ruddock, Deputy Head of the Department of Assessment and Measurement, National Foundation for Educational Research in England and Wales, responded to my query with respect to the national tests for 7, 11, and 14 year-olds. He wrote, "A consistent approach to cutoffs has not yet been reached" (personal communication, August 1994). Ruddock reported that a procedure based on hierarchies of levels is used.

> In mathematics for 1994, Angoff's procedure was used to set cutoffs, but the results from this were slightly modified after inspection. In other subjects, and as a supplementary procedure in mathematics, algorithmic approaches are used. These involve setting cutoffs so that a pupil awarded level n must have achieved substantial success on level n items and cannot achieve level n by mostly answering easier items at levels n-1.
>
> The process is carried out by the test developers working with the professional officers of the School Curriculum and Assessment Authority, the government agency which commissions the tests. It is not in the public domain and not publicly documented. (personal communication, August 1994)

Jim Tognolini, Director of the Educational Testing Centre, the University of New South Wales, responded to my request for information.

> In Australia there are no national examinations per se. The different States have basic skills testing. The results of this testing are usually reported in terms of bands which are characterized by descriptive curriculum statements. Cut-off scores for the bands are determined by committees looking at past years' standards and making informed but subjective decisions about the boundaries for the current test. This process does not engender any real interest at the political level nor at the media level. (personal communication, August 1994)

David M. Elliot, Director of Assessment, Scottish Examination Board, also responded to my request. He described an elaborate system of standard setting in use for The Standard Grade Test, which is offered at three levels and is taken by 97 percent of the students in the fourth year of secondary school in Scotland. The Angoff and Nedelsky methods had been tried and were found wanting because they disagreed with each other and because there was great variability across raters within method. Cutoff score ranges were set (for example 70-80% correct). The ranges were for guidance only, and cut scores could be set outside the range.

Equipercentile equating is used to assure that the cut points on the three levels are sensible. That is, it should not be harder to pass a lower-level test than an upper-level test. In addition, "comparability indices" are checked across the tests in 34 subjects.

Elliot indicated that a method called the "expert group" approach was also used.

> A breakthrough came from consideration of the *expert group* approach. In Scotland we have for decades asked teachers to estimate the performance of their students on our tests; if the student fails to perform as expected an *appeal* can be submitted, supported by classroom evidence. . . . As these estimates were already available, it was a simple matter to compute the scores on the tests which would produce the same proportion of students on each grade. . . . The teachers base their judgments on the Grade Related Criteria and outcomes of class tests. (personal communication, August 1994)

The Principal Examiner provides a qualitative view based on reports from the various markers concerning the perceived difficulty of the test and how well students performed compared to previous years. All the data are brought to a high-level committee, and after consideration of the evidence, tentative cutoff scores are set. The effects of the tentative cutoff scores are then determined and adjusted if necessary.

Elliot reported, "With regard to public acceptance, it has to be said that there is a different tradition in this country regarding disclosure from that in the U.S. The cut-off scores are confidential" (personal communication, August 1994). He also noted that the public tends to accept the board's judgment and that there is little or no controversy concerning the cut-off scores.

What can be learned from the ways that standards are set in other countries? The major lesson appears to be that the use of normative information as a moderating influence in the adjustment of standards helps maintain consistent passing rates and aids in public acceptance of the results.

THE AGE OF REALISTIC ACCEPTANCE

Calling the present the Age of Realistic Acceptance may represent the triumph of hope over experience, but enough information should have been accumulated by now to bring rational measurement professionals to a state of realistic acceptance of the basic characteristics of standard setting, even though there may be disagreement about the nuances.

What have we learned about standard setting from the time of Abraham to the most recent efforts to apply standard-setting methodologies to computerized adaptive tests? What has history taught us?

We have learned that all methods of setting standards depend on judgment.

This is probably the single greatest area of agreement in the history of setting standards. Sometimes the judgments are applied to people or products, and sometimes the judgments are applied to test questions, but *no* method of setting standards is completely free of judgment. In reviewing the literature, I did not find a single author who disagreed with the fact that standard setting requires judgment. In fact, many authors stated quite explicitly and eloquently that judgment is required.

We do not shirk from judgment in other aspects of life. We set standards because they are often useful and sometimes necessary. As Livingston and Zieky (1982) noted, "There are standards for many kinds of things, including the purity food products, the effectiveness of fire extinguishers, and the cleanliness of auto exhaust fumes" (p. 10). Those standards are all "arbitrary," yet they are beneficial. In effect, the people who continue to set standards admit that Glass (1978) was correct about the arbitrary nature of standards, but they reject the notion that being arbitrary, in the sense of being judgmental, is a fatal flaw.

We have learned that standard setting will lead to errors of classification.

Standards almost always impose an artificial dichotomy on an essentially continuous distribution of knowledge, skill, and ability. Few tests in the world are able to distinguish reliably between examinees at adjacent score levels. Yet the imposition of a standard will cause examinees at some score level to be classified in one way and examinees at an adjacent score level to be classified in a different way.

Implicitly or explicitly, the standard is a manifestation of the value judgments inherent in determining if it is preferable to pass an examinee who deserves to fail, or if it is preferable to fail an examinee who deserves to pass. Because no test can be perfectly reliable or perfectly valid, such mistakes in classification will be made if people are to be split into passing and failing groups.

When setting standards, it is impossible to reduce one of the types of mistakes in classification without increasing the other type. Standards that are set relatively high will reduce the chances of passing people who should fail, but such standards will increase the chances of failing people who should pass. Standards that are set relatively low will reduce the chances of failing people who should pass, but will increase the chances of passing people who should fail. If the test is improved, the errors of classification can be reduced, but they will never be reduced to zero.

We have learned that standard setting is controversial and will necessarily remain so.

Even though the methodology for setting standards on conventional tests has evolved over a number of years to the point at which there are several methods or procedures that are in general use, standard setting remains highly controversial. The reason for this controversy is that a standard is *not* a statement of some psychometrically derived truth. A standard is, rather, the embodiment of a set of value judgments. As long as different people hold different values, standard setting will remain a matter of controversy.

Because different people hold different values about which type of mistake in classification is worse, a standard that is seen by some as absurdly low, an embarrassment to educators or the profession, and a danger to the public will be seen by others as unfairly high, a barrier to aspiring examinees, and the cause of avoidable adverse impact. Because the standard depends so greatly on value judgments, it is difficult to defend the standard against critics who have very different values than those of the people involved in setting the standard.

We have learned that there is NO "true" standard that the application of the right method, in the right way, with enough people, will find.

Jaeger (1989) said that "Much early work on standard setting was based on the often unstated assumption that determination of a test standard parallels estimation of a population parameter" (p. 492). The statistical training of measurement professionals involved in setting standards seems to lead many of them to believe that better studies using larger samples or more refined methodology will reduce the sampling error and bring them ever closer to knowledge of the "true" standard.

BEST COPY AVAILABLE

It is indeed possible to conceive of the results of a standard-setting study as an estimate of a population parameter. That parameter, however, is not the true standard. It is simply the result that would have been obtained had it been possible to run that particular study without any sampling error. Some other equally justifiable methodology or a sample from some other equally acceptable population of judges would have resulted in an estimate of a different parameter.

It is a fallacy that a true standard exists and that it could be found, or gotten very close to, if only we tried hard enough. Standards are constructed, not discovered. Jaeger (1989) wrote convincingly of the fact that the true standard is an illusion. "If competence is a continuous variable, there is clearly no point on the continuum that would separate students into the competent and the incompetent. A right answer does not exist, except perhaps in the minds of those providing the judgments" (p. 492).

We have learned that there is no universally accepted "right" way to set standards.

There are now many compendiums of methods of setting standards. Some methods have been used many times in many situations and have become widely accepted. They are the methods summarized and evaluated time and time again in the references cited above: Angoff's, Ebel's, Nedelsky's, and Jaeger's methods, the borderline-group and contrasting-groups methods, and their various documented modifications.

Even those methods, however, are not universally accepted in every context. A National Academy of Education (Shepard, Glaser, Linn, & Bohrnstedt, 1993) report, for example, cited problems with the use of the Angoff method to set standards on the National Assessment of Educational Progress.

> The Angoff method approach and other item-judgment methods are fundamentally flawed. Minor improvements . . . cannot overcome the nearly impossible cognitive task of estimating the probability that a hypothetical student at the boundary of a given achievement level will get a particular item correct.[8] (p. xxiv)

Furthermore, there is no document equivalent in force and stature to the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985) that describes standard-setting methodologies in sufficient operational detail to allow practitioners to feel free from criticism because they "did it by the book." There is no single method or way of applying that method that is invulnerable to attack from critics.

---

[8] I don't have any research to verify this, but my experience in standard setting leads me to believe that judges involved in the major item-judgment methods of setting standards are not really making estimates of the performance of some hypothetical group of minimally competent examinees.

I think the actual process is closer to one in which a judge looks at an item and decides that he or she would not be willing to call an examinee minimally competent unless the examinee had a chance of at least *x* of getting the item right. I believe that judges are not making sloppy estimates of a probability. They are directly expressing their own values about what level of performance they would consider good enough to call minimally competent. (There is a good dissertation topic lurking here for some graduate student.)

There are so many variables involved in an operational standard-setting and so many opinions and theories about each variable that some aspects of any study will always be open to attack if critics are unhappy with the results and wish to undermine the study. A standard-setting study that is planned and carried out by recognized experts can be severely criticized by other recognized experts and defended by yet other recognized experts. (See, for example, U.S. General Accounting Office, 1993a.)

We have learned that there are no purely absolute standards.

All standards have a normative component because judgments of what constitutes minimal competence are necessarily affected by what people can actually do with respect to the knowledge, skills, and abilities measured by the test. No judge would set the minimum passing score on a typing test at 1,800 words per minute, for example, nor would any judge set it at 2 words per minute.

Because standards are necessarily affected by whatever normative information the judges happened to have internalized, it makes perfect sense to me to ensure that the judges involved in setting standards have access to accurate, relevant, normative data. If there is a desire to get an "uncontaminated" absolute standard, I strongly suggest using an iterative approach and providing normative information to the judges at some later stage of the process.

We have learned that standard setting should be part of a process, not an isolated act.

Because of the clearly judgmental and value-loaded nature of setting standards, running the standard-setting study should be only one of the steps in the process of setting standards. The entire standard-setting process should include the steps of planning the process, informing those affected, gathering data, selecting a standard, and evaluating the effects of the standard.

The decision makers who are responsible for actually determining the standards to be used should be intimately involved in planning the standard-setting process. The people who will be affected by the standards (or their parents or guardians) should be informed about the relevant issues. If possible, representative groups of stakeholders should be invited to participate in the process in some way.

Measurement experts should be involved in pointing out alternatives in standard setting, explaining the expected advantages and disadvantages of each alternative, offering recommendations if sufficient data and experience exist to justify them, and implementing the alternatives that the decision makers select.

Because judgment is essential and because judgments differ, it is crucial to involve judges that are acceptable to the constituencies that have a stake in the outcome of the standard-setting process.

Once the data have been collected from the judges, regardless of the method used, the actual standard may be set in several ways. The number that comes out of the standard-setting study is not sacrosanct. Had a different method been used, or if the method had been applied in a different way, or if different judges had been involved, the result of the standard-setting study would very probably have been different. Decision makers should use the result of the study as one very important piece of information, but it is reasonable for them to adjust the standard in light of their perceptions of the relative harm caused by the two types of classification errors in the use of the test for its intended purpose.

Finally, the effects of using the standards operationally should be monitored. Are they working as intended?

We have learned that we must be ready to modify standards on the basis of experience.

Because of the judgmental nature of the standard-setting process and the lack of universally accepted methods of gathering those judgments, there is always a possibility that something might have gone wrong. Even if the entire process has been implemented flawlessly, there may be unanticipated negative consequences. It is important to evaluate the results of applying the standard that was set and to be ready to modify that standard if experience demonstrates that it is resulting in inappropriate consequences that outweigh the appropriate ones.

We must keep in mind, however, that there is no objective proof of the validity of a standard. People may certainly disagree in assigning costs to what may be perceived as inappropriate consequences and benefits to what may be perceived as appropriate consequences. They may even disagree as to whether a particular outcome is appropriate or not.

Whether a standard is to be considered valid or not is a matter of values. The fact that 98% of the people believe a standard is working appropriately, based on the evidence that has been collected, does not mean that the remaining 2% are wrong. They may have very different values concerning the relative harm caused by the two types of errors of classification. For them the standard may actually not be valid.

A CLOSING WORD

What is the best way to summarize what has been learned so far about setting standards? Ebel's (1972) wise comments seem highly appropriate, and they should be required reading for all who become engaged in the process of setting standards.

> Anyone who expects to discover the "real" passing score by any of these approaches, or any other approach, is doomed to disappointment, for a "real" passing score does not exist to be discovered. All any examining authority that must set passing scores can hope for, and all any of their examinees can ask, is that the basis for defining the passing score be defined clearly, and that the definition be as rational as possible. (p. 496)

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement, 36*, 45-50.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education, 45,* 4-9.

Berk, R. A. (Ed.). (1980). *Criterion-referenced measurement: The state of the art.* Baltimore: Johns Hopkins University Press.

Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Meesurement, 21*, 147-152.

Block, J. H. (1978). Standards and criteria: A response. *Journal of Educational Measurement, 15,* 291-295.

Broadfoot, P. (1992). Assessment developments in French education. *Educational Review, 44,* 309-326.

Buck, L. S. (1977). *Guide to the setting of appropriate cutting scores for written tests: A summary of the concerns and procedures.* Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center.

Burton, N. W. (1978). Societal standards. *Journal of Educational Measurement, 15,* 263-271.

Cohen, R. J., Montague, P., Nathanson, L. S., & Swerdlik, M. (1988). *Psychological testing.* Mountain View, CA: Mayfield Publishing.

Conaway, L. E. (1979). Setting standards in competency-based education: Some current practices and concerns. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based measurement* (pp. 72-88). Washington, DC: National Council on Measurement in Education.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart, and Winston.

Department of Education and Science, Welsh Office. (1982). *Examinations at 16-plus: A statement of policy.* London: Her Majesty's Stationery Office.

De Gruijter, D. N. M. (1985). Compromise models for establishing examination standards. *Journal of Educational Measurement, 22,* 263-269.

Ebel, R. L. (1965). *Measuring educational achievement.* Englewood Cliffs, NJ: Prentice-Hall.

Ebel, R. L. (1972). *Essentials of educational measurement* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Eckstein, M. A., & Noah, H. (1993). *Secondary school examinations: International perspectives on policies and practice.* New Haven, CT: Yale University Press.

Faggen, J. (1994). *Setting standards for constructed response tests: An overview.* Princeton, NJ: Educational Testing Service.

Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.), *Educational measurement* (1st ed., pp. 695-763). Washington, DC: American Council on Education.

Geisinger, K. F. (1991). Using standard-setting data to establish cutoff scores. *Educational Measurement: Issues and Practice, 10,* 17-22.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18,* 519-521.

Glaser, R., & Klaus, D. J. (1962). Proficiency measurement: Assessing human performance. In R. M. Gagne (Ed.), *Psychological principles in systems development*. New York: Holt, Rinehart, and Winston.

Glaser, R., & Nitko, A. J. (1971). Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 625-670). Washington, DC: American Council on Education.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15,* 237-261.

Greaney, V., & Kellaghan, T. (1979). In F. M. Ottobre (Ed.), *Criteria for awarding school leaving certificates. An international discussion.* Oxford: Pergamon.

Green, B. F. (1981). A primer of testing. *American Psychologist, 36,* 1001-1011.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley. (Reprinted, 1987, Hillsdale, NJ: Erlbaum).

Hambleton, R. K. (1978). On the use of cut-off scores with criterion referenced tests in instructional settings. *Journal of Educational Measurement, 15,* 277-290.

Hambleton. R. K. (1980). Test score validity and standard-setting methods. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art* (pp. 80-123). Baltimore: Johns Hopkins University Press.

Hambleton, R. K., & Novick, M. R. (1972). *Toward an integration of theory and method for criterion-referenced tests.* Iowa City, IA: American College Testing.

Hambleton, R. K., & Powell, S. (1981, April). *Standards for standard setters.* Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Hambleton, R. K., & Rogers, J. H. (1991). Advances in criterion-referenced measurement. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 3-44). Boston: Kluwer Academic.

Hills, J. R. (1971). Use of measurement in selection and placement. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 680-732). Washington, DC: American Council on Education.

Jaeger, R. M. (1978). *A proposal for setting a standard on the North Carolina high school proficiency test.* Paper presented at the spring meeting of the North Carolina Association for Research in Education, Chapel Hill.

Jaeger, R. M. (1979). Measurement consequences of selected standard-setting models. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based measurement* (pp. 72-88). Washington, DC: National Council on Measurement in Education.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4,* 461-475.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). New York: American Council on Education/Macmillan.

Jaeger, R. M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice, 10,* 3-6, 10.

Jaeger, R. M. (1994, April). *Setting performance standards through two-stage judgmental policy capturing.* Paper presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, New Orleans.

Lindquist, E. F. (Ed.). (1951). *Educational measurement* (1st ed.). Washington, DC: American Council on Education.

Linn, R. L. (Ed.). (1989). *Educational measurement* (3rd ed.). New York: American Council on Education/Macmillan.

Livingston, S. A. (1976). *Choosing minimum passing scores by stochastic approximation techniques.* Princeton, NJ: Educational Testing Service.

Livingston, S. A. (1980). Comments on criterion-referenced testing. *Applied Psychological Measurement, 4,* 575-581.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Madaus, G. F., & Kellaghan, T. (1991). *Examination systems in the European Community: Implications for a national examination system in the United States.* Springfield, VA: U.S. Department of Commerce, National Technical Information Center.

Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. *Review of Educational Research, 46,* 133-158.

Millman, J. (1973). Passing scores and test lengths for domain-referenced measures. *Review of Educational Research, 43,* 205-217.

Mills, C. N., Melican, G. J., & Ahluwalia, N. T. (1991). Defining minimal competence. *Educational Measurement: Issues and Practice, 10,* 7-10, 14.

Monroe, W. S., DeVoss, A. M., & Kelly, F. J. (1924). *Educational tests and measurements.* Cambridge, MA: Riverside Press.

National Academy of Education (Shepard, Glaser, Linn, & Bohrnstedt, 1993). *Setting performance standards for student achievement.* Washington, DC: Author.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Nitko, A. J. (1983). *Educational tests and measurement: An introduction.* New York: Harcourt Brace Jovanovich.

Peterson, J. J. (1983). *The Iowa testing programs.* Iowa City: University of Iowa Press.

Plake, B. S., Melican, G. J., & Mills, C. N. (1991). Factors influencing intrajudge consistency during standard-setting. *Educational Measurement: Issues and Practice, 10,* 15-16, 22, 25-26.

Popham, W. J. (1978). As always, provocative. *Journal of Educational Measurement, 15,* 297-300.

Reddaway, J. L. (1988). Examinations for university selection in England. In S. P. Heyneman & I. Fagerlind (Eds.), *University examinations and standardized testing: Principles, experience, and policy options* (pp. 26-34). Washington, DC: International Bank for Reconstruction and Development.

Reid, J. B. (1991). Training judges to generate standard-setting data. *Educational Measurement: Issues and Practice, 10,* 11-14.

Scriven, M. (1978). How to anchor standards. *Journal of Educational Measurement, 15,* 273-275.

Shepard, L. A. (1976, April). *Setting standards and living with them.* Paper Presented at the annual meeting of the American Educational Research Association, San Francisco.

Shepard, L. A. (1979). Setting standards. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based measurement* (pp. 72-88). Washington, DC: National Council on Measurement in Education.

Shepard, L. A. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4,* 447-467.

Shimberg, B. (1982). *Occupational licensing: A public perspective.* Princeton, NJ: Educational Testing Service.

Stocking, M. (personal communication 1992).

Sutherland, G. (1984). *Ability, merit and measurement: Mental testing and English education, 1880-1940.* Oxford: Clarendon Press.

Thorndike, R. L. (Ed.). (1971). *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Thorndike, R. L., & Hagen, E. P. (1977). *Measurement and evaluation in psychology and education* (4th ed.). New York: Wiley.

Thorndike, R. M., & Lohman, D. F. (1990). *A century of ability testing.* Chicago: Riverside Publishing.

U.S. Congress, Office of Technology Assessment. (1992, February). *Testing in American schools: Asking the right questions* (1992-297-934Q13). Washington, DC: U.S. Government Printing Office.

U.S. Department of Justice. (1976). Federal executive agency guidelines on employee selection procedures. *Fed. Reg., 41,* 51733-51759.

U.S. General Accounting Office. (1993a). *Educational achievement standards: NAGB's approach yields misleading interpretations* (Rep. No. GAO/PEMD-93-12). Washington, DC: Author.

U.S. General Accounting Office. (1993b). *Educational testing: The Canadian experience with standards, examinations, and assessments.* Washington, DC: Author.

van der Linden, W. J. (1980). Decision models for use with criterion-referenced tests. *Applied Psychological Measurement, 4,* 469-492.

van Krieken, R. (1987). Marking and examinations in the Netherlands. *Studies in Educational Evaluation, 13,* pp. 35-42.

Zieky, M. J. (1973). *Methods of setting standards for criterion referenced item sets*. Princeton, NJ: Educational Testing Service.

Zieky, M. J. (1987, November). *Methods of setting standards of performance on criterion referenced tests.* Paper presented at the 13th International Conference of the International Association for Educational Assessment, Bangkok.

Zieky, M. J., & Livingston, S. A. (1977). *Manual for setting standards on the Basic Skills Assessment tests*. Princeton, NJ: Educational Testing Service.

Standards for Reporting the Educational Achievement of Groups

Samuel A. Livingston

Senior Measurement Statistician, Educational Testing Service, New Jersey

ABSTRACT

Standard-setting studies translate the judges' conceptual standards into an implied operational standard for making decisions or reporting group percentages. The operational standards implied by item-judgment studies and by borderline-group studies tend to produce a biased estimate of the percentage of students who meet the judges' conceptual standard. The operational standard that minimizes the number of decision errors tends to produce an estimate that is biased in the opposite direction. However, these two standards can be combined to produce an operational standard that, under certain plausible assumptions, produces an unbiased estimate.

A standard is an answer to the question, "How good is good enough?" Standard setting is the process of answering this question in such a way that the answer can be used for a particular purpose. In education, the purpose of setting a standard is often to make decisions about individuals. Should this student be admitted to an advanced course? Should this student be required to take a remedial course? Should this student be given credit for a course taken or for a course not taken? Should this teacher be certified to teach a particular subject? And so on. But sometimes the purpose of setting a standard is to report information about groups. Setting a standard makes it possible to report information about students' educational achievements in terms of a simple, familiar, and easily understood statistic--a percentage. When people read about the percentage of students in a group who meet the standard, they may be confused about which students are included in the group or about what it means to meet the standard, but they are not confused about what the percentage is.

CONCEPTUAL STANDARDS AND OPERATIONAL STANDARDS

One important distinction that people need to make when they think about setting standards in education is between *conceptual* standards and *operational* standards. A conceptual standard is the concept in a person's mind that enables that person to decide whether something is or is not good enough. An operational standard is a rule for deciding whether something is or is not good enough. Conceptual standards are implicit; operational standards are explicit. Conceptual standards are subjective; operational standards are objective. Conceptual standards are generally not quantified; operational standards are almost always quantified. A person's conceptual standard refers to all those proficiencies that the person considers relevant for classifying the student--and only those proficiencies. An operational standard refers to only those proficiencies that are actually measured, usually by some kind of test. Therefore, operational standards often take the form of "cut scores."

The distinction between a conceptual standard and an operational standard is not simply a distinction between a "true-score" standard and an "observed-score" standard. The true score of classical-test theory is a long-term average over repeated testing, that is, the average of the scores the student would earn on all possible versions of the test, taken under all permissible conditions, and so forth. The true score and the observed score measure exactly the same proficiencies, with the exception of an additional random "error" component in the observed score. In contrast, the latent variable underlying a person's conceptual standard may not include exactly the same proficiencies measured by the test, even if that person is the one who selected the test. There should be a large overlap, but each set may include some proficiencies not included in the other. I will refer to the latent variable underlying a person's conceptual standard as that person's "concept of proficiency," to emphasize its subjective nature.

STANDARD-SETTING STUDIES AND STANDARD-SETTING DECISIONS

A standard-setting study is a way of translating conceptual standards into an operational standard. The conceptual standards involved are those of the people who serve as judges in the study. The study translates their conceptual standards onto the test-score scale. The judges' concept of proficiency may or may not be the same as that of the people who created the test. If the judges' concept of proficiency is highly similar to that of the test makers, the judgments in the study are likely to imply a clear operational standard. If the judges' concept of proficiency is quite different from that of the test makers, the judgments may not clearly imply an operational standard. This possibility raises an important issue for those who conduct standard-setting studies: Should they insist that the judges adopt the test makers' point of view as to which proficiencies are relevant? The answer to this question may depend on factors specific to the situation: Who are the judges, how were they selected, how will the results of the study be used, and so on.

A standard-setting decision (the choice of an operational standard) is a statement of educational policy. An operational standard should be chosen by a person or group with the authority and responsibility to make a policy decision. A standard-setting study provides one kind of information for making this decision. It identifies the operational standards that are implied by the conceptual standards of selected individuals the judges selected for the study. The standard setters may have very good reasons for taking other kinds of information into account. Limited resources may put a restriction on the range of operational standards that can actually be implemented. It may be important to avoid a large imbalance between different groups of students in the awarding of some educational benefit. These other kinds of information may imply a need for an operational standard different from the one implied by the standard-setting study. This kind of conflict is real, and it has nothing to do with measurement error in the test scores. The conflict would exist even if the test scores were perfectly reliable. Therefore, the way to resolve it is not to invoke the standard error of measurement, as has sometimes been done. The people responsible for choosing the operational standard must treat this type of conflict as they would any other conflict between competing considerations in a policy decision. The issue is one of educational or social policy, not of psychometrics.

TYPES OF STANDARD-SETTING STUDIES

Standard-setting studies in education can be classified into two general categories on the basis of the type of judgments they involve. Item-judgment studies involve judgments about the questions, tasks, or problems on the test. Person-judgment studies involve judgments about actual test takers.

Typically, item-judgment studies require the judges to say how a particular kind of test taker would respond to each question on the test. This test taker is one whose proficiency, as conceptualized by the judges in the study, is on the borderline between adequate and inadequate. Nedelsky (1954) referred to this test taker as the "F-D student." Angoff (1971) used the term "minimally acceptable person." In this paper, I will use the term "borderline student."[1]

There are two general types of person-judgment studies. In a borderline-group study, the judges identify those test takers who are borderline students with respect to the proficiencies that are relevant for the decision or classification to be made. In a contrasting-groups study, the judges classify test takers (or samples of their work) as representing adequate or inadequate proficiency.

BIAS IN ESTIMATING THE PERCENTAGE OF STUDENTS WHO MEET THE STANDARD

Borderline-group studies are based on the same reasoning as item-judgment studies. That reasoning can be expressed this way:

> Any student more proficient than the borderline student must be above the standard; any student less proficient than the borderline student must be below the standard. Therefore, any student with a test score higher than the score of a typical borderline student should be classified as meeting the standard. Any student with a lower test score should be classified as not meeting the standard. It follows that the standard, expressed on the test-score scale, is the test score of a typical borderline student.

This reasoning seems clear and compelling, but it leads to a problem if the operational standard implied by a borderline-group study is used as the basis for reporting group percentages. Figure 1 illustrates this problem. Imagine that it were possible not only to test each student in the population, but also to know the student's standing with respect to the judges' concept of proficiency. If you could construct a scatterplot of these two variables, the cluster of points would tend to form an approximately elliptical shape. That is what the ellipse in Figure 1 represents. The horizontal line through the ellipse represents the judges' conceptual standard. In this example, most of the students are above the judges' conceptual standard. The borderline group is the group of students whose proficiency is exactly at the judges' conceptual standard. This group is represented by the solid portion of the horizontal line, inside the ellipse. The midpoint of this line represents the test score of a typical borderline student. The vertical line through this point represents the operational standard implied by a borderline-group study (and also by an item-judgment study that is working as intended).

In Figure 1, the area inside the ellipse and below the horizontal line represents the students whose proficiency is below the judges' conceptual standard. The area inside the ellipse and to the left of the vertical line represents the students whose test scores are below the operational standard implied by the study. Notice that these two areas are not the same size. The percentage of students with test

---

[1] I am skeptical as to whether judges in a standard-setting study can actually conceptualize a borderline student and determine such a student's expected performance on a test item. However, my colleague Michael Zieky has provided a rationale for item-judgment studies that seems to me to be more realistic. He says that what the judges in these studies are really telling us is how well a test taker would have to be able to perform on the item in order for the judge to consider the test taker minimally competent. If a judge in a Nedelsky study eliminates options A and C, the judge is saying, in effect, "I would not consider a test taker even minimally competent unless that individual knows that options A and C are wrong." If a judge in an Angoff study specifies a correct-answer probability of .80, the judge is saying, in effect, "I would not consider a test taker even minimally competent unless that individual has at least an 80 percent probability of answering this item correctly."

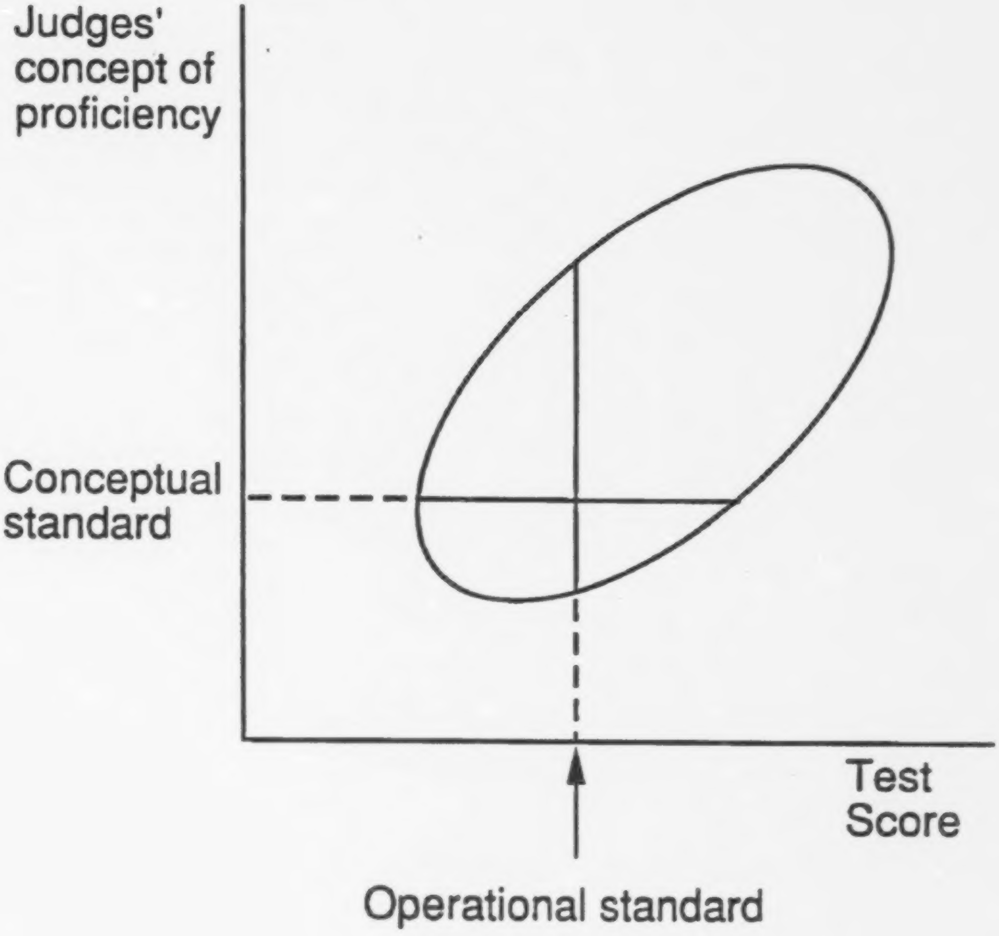

Judges' concept of proficiency
Conceptual standard
Test Score
Operational standard

Figure 1. Operational standard implied by a borderline-group or item-judgment study.

scores below the operational standard implied by the study will be a *biased estimate* of the percentage of students whose proficiency (as conceptualized by the judges in the study) is below the judges' conceptual standard. The same will be true of the operational standard implied by an item-judgment study, if the study works as it is supposed to, because that standard represents the test score of a typical borderline student.

The cause of the bias is a regression effect, "regression to the mean." In the example of Figure 1, the borderline students are below average. These students are identified as borderline on the basis of the judges' concept of proficiency, which is not perfectly correlated with the test scores. Therefore, the test score of a typical borderline student will be regressed toward the mean of the group from which the borderline students were selected. If you use this test score as the operational standard, that standard will be too close to the mean, if your purpose is to determine what percentage of the students fall below the judges' conceptual standard.

MINIMIZING DECISION ERRORS

A contrasting-groups study provides the data necessary for an analysis based on the logic of decision theory. When the test scores are used to make decisions about individual students, two kinds of decision errors are possible. A student whose proficiency is below the judges' conceptual standard can have a test score above the operational standard; this type of error is sometimes called a "false positive." And a student whose proficiency is above the judges' conceptual standard can have a test score below the operational standard; this type of error is sometimes called a "false negative." Raising the operational standard will reduce the number of false positives at the cost of increasing the number of false negatives. Lowering the operational standard will have the opposite effect. Decision theory tells you to choose the operational standard that will minimize the total harm from the false positives and the false negatives. If the two types of errors are equally harmful, decision theory tells you to choose the operational standard that minimizes the total number of decision errors.

Notice in Figure 1 that the operational standard implied by a borderline-group study will not minimize the total number of decision errors. Most of the students with test scores just below the operational standard will be above the judges' conceptual standard in proficiency. You could reduce the number of decision errors by lowering the operational standard to classify these students into the higher group. To minimize the total number of decision errors, you should lower it to the point indicated in Figure 2, the point at which half the students have proficiency levels above the judges' conceptual standard and half the students have proficiency levels below that standard. Again, the area inside the ellipse and below the horizontal line represents the students whose proficiency is below the judges' conceptual standard. The area inside the ellipse and to the left of the vertical line represents the students whose test scores are below the operational standard that minimizes the number of decision errors. Again, the two areas are not the same size. The percentage of students with test scores below the operational standard implied by the study will be a *biased estimate* of the percentage of students whose proficiency is below the judges' conceptual standard--but this time the bias is in the opposite direction.

Again, the cause of the bias is a regression effect, regression to the mean. This time, you are classifying students on the basis of their test scores. No matter what test score you choose, the proficiency levels of the students with that score will tend to be regressed toward the mean. You are trying to find the test score at which, according to the judges' concept of proficiency, half the students are above the judges' conceptual standard and half the students are below that standard. Therefore, you will have to choose a test score farther from the mean than the judges' conceptual standard is.

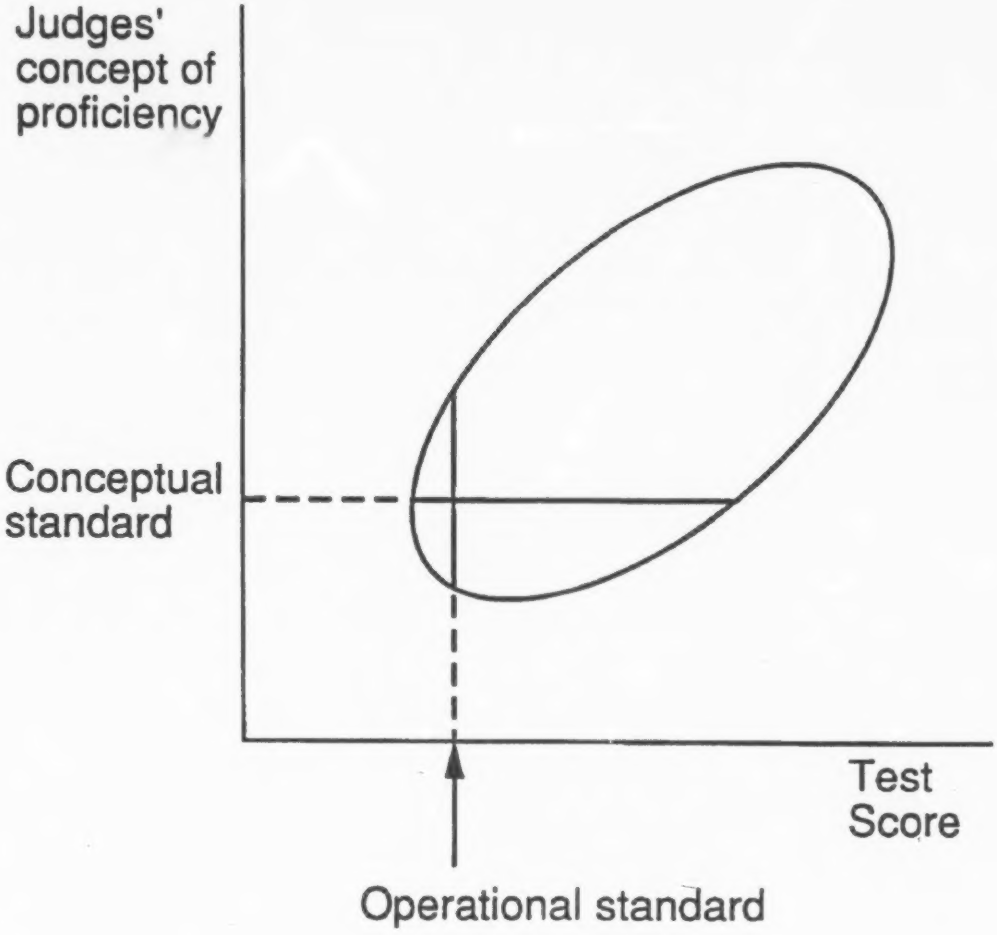

Judges' concept of proficiency
Conceptual standard
Test Score
Operational standard

Figure 2. Operational standard chosen to minimize decision errors.

An extreme example occurs when even the lowest scoring students have a better-than-even chance of being judged proficient. (Fortunately, this situation does not occur often.) In this case, you can minimize the number of *individual* decision errors by classifying every student as proficient, even though you know that some of them are not. The operational standard that minimizes the number of decision errors will be below the test score of the lowest scoring student. Clearly, the percentage of students with test scores below this operational standard (i.e., 0%) is a biased estimate of the percentage whose proficiency is below the judges' conceptual standard.

HOW LARGE IS THE BIAS?

The bias in estimating the percentage of students below the judges' conceptual standard is greatest when the actual percentage is very small or very large. When this percentage is close to 50%, the bias is much smaller, as illustrated in Figure 3. When the proficiency of the average student is exactly at the judges' conceptual standard, the bias disappears.

The bias also depends on the relationship between the students' test scores and their proficiency as perceived by the judges. The stronger this relationship the less the bias. Figure 4 illustrates the case in which the students' test scores are very highly correlated with their proficiency as perceived by the judges. In this case, the bias is quite small. If the correlation were perfect, there would be no bias at all.

To determine the size of the bias in numerical terms, it is necessary to make some assumptions. Table 1 shows the results of some theoretical calculations based on a model in which the students' test scores and their actual proficiency levels are assumed to have a bivariate normal distribution.[2] This model is not intended as an accurate picture of reality, but it should provide at least a rough approximation to what happens in a real standard-setting situation. When the correlation between the test scores and the judges' concept of proficiency is .90, the bias is fairly small, about 1 to 3 percentage points. But when the correlation falls to .70, the bias is about 6 to 9 percentage points (unless the judges' conceptual standard divides the group nearly in half). With a correlation of .70, if 10% of the students are below the judges' conceptual standard, the operational standard implied by a borderline-group study will show 18% of the students below standard, instead of 10%. (An item-judgment study will have the same effect if it works exactly as intended.) Under the same circumstances, the operational standard that minimizes the number of decision errors will show only 3% of the students below standard.

AN UNBIASED ESTIMATE

So far, I have discussed two ways of using a standard-setting study to identify an operational standard. If these operational standards are used to estimate the percentage of students whose proficiency is below the judges' conceptual standard, the resulting estimates are biased in opposite directions. Can you combine them to get an estimate that is less biased, or even unbiased? The answer is yes. For the estimate to be unbiased, the regressions in the scatterplot shown in Figures 1 and 2 must be linear and the conditional distributions must be symmetric. The judges' concept of proficiency is not observable, so you can never know how close to reality these conditions are. However, it seems reasonable to

---

[2] The derivations of the equations for the operational standards are shown in the Appendix. My thanks to Neal Thomas for his help with these derivations and to Jim Ferris for his help in producing the numerical estimates.

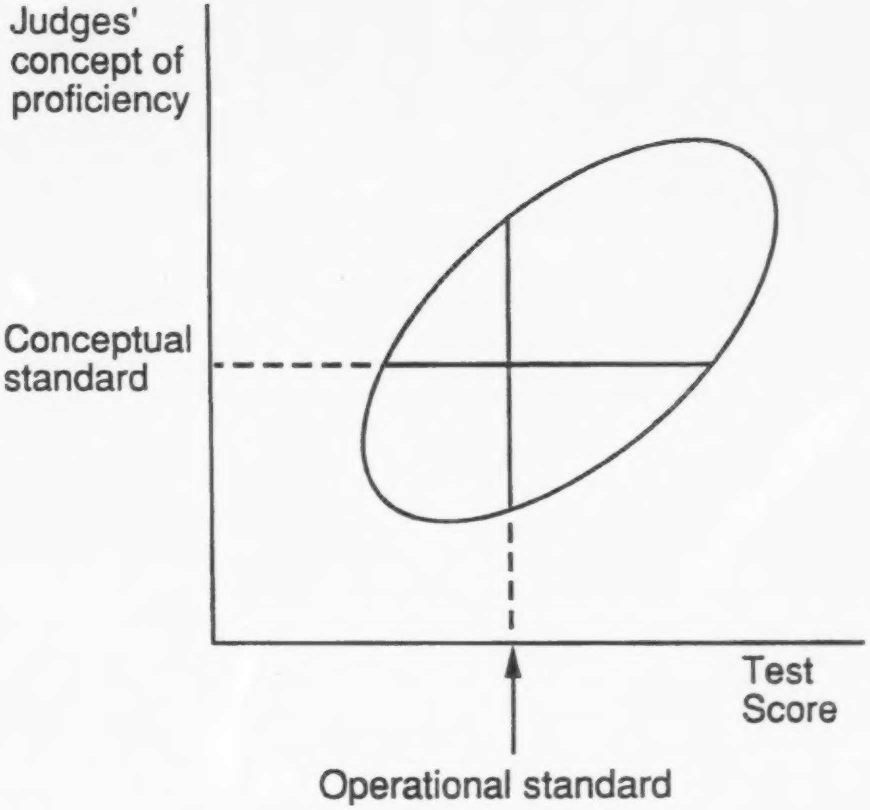

Judges' concept of proficiency
Conceptual standard
Test Score
Operational standard

Figure 3. Conceptual standard near middle of group. Operational standard chosen to minimize decision error.

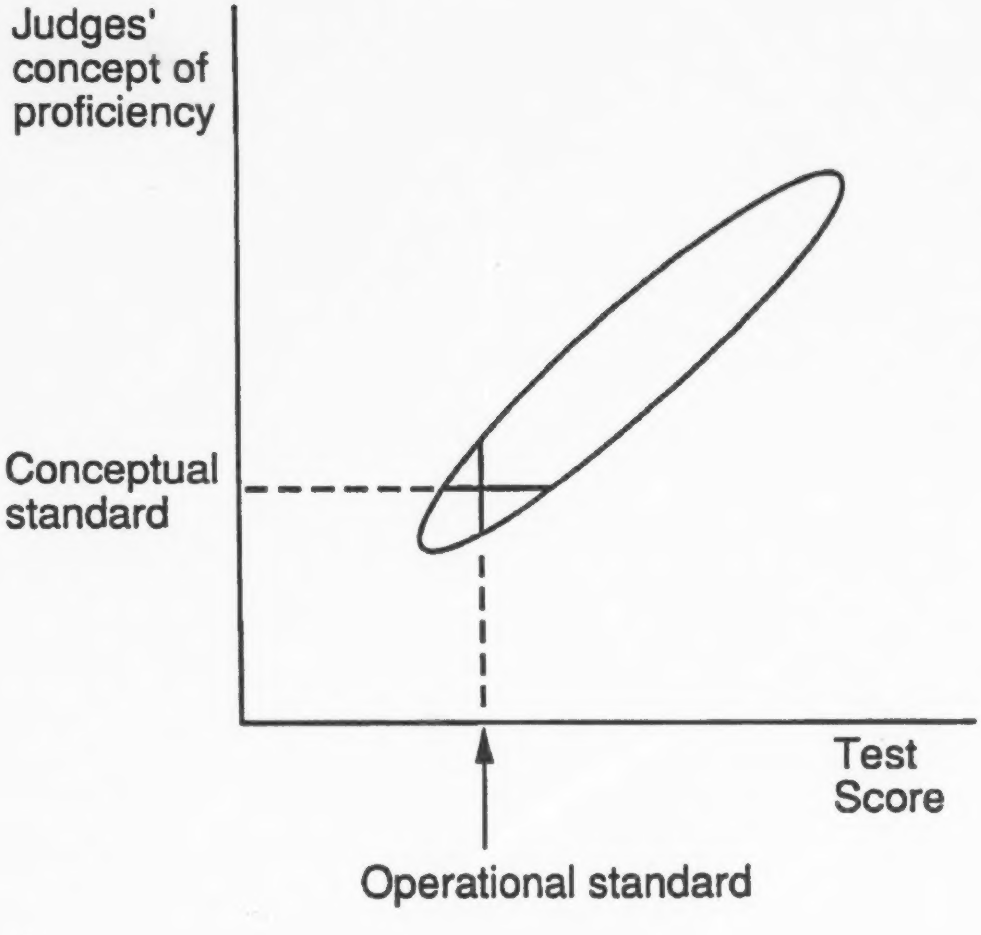

Judges' concept of proficiency
Conceptual standard
Test Score
Operational standard

Figure 4. Variables strongly correlated. Operational standard chosen to minimize decision errors.

Table I

Percentages of Students Below Conceptual and Operational Standards: Theoretical Calculation Based on Bivariate Normal Model

Correlation of test scores with judges' concept of proficiency	Conceptual standard	Percentage of students below... Operational standard: Borderline group or item-judgment	Percentage of students below... Operational standard: Minimizing decision errors
.90	40	41	39
	30	32	28
	20	22	17
	10	12	7
	5	7	3
.70	40	43	36
	30	36	23
	20	28	11
	10	18	3
	5	12	1

assume that they are approximately true, especially if the distribution of test scores is approximately symmetric. (If it is not, you can transform the test scores onto a scale that will give them an approximately symmetric distribution.)

To estimate the percentage you are looking for, it is convenient to express the judges' conceptual standard and the operational standards as z-scores (i.e., in terms of standard deviations from the mean). If z_1 represents the operational standard implied by a borderline-group study (or by an item-judgment study that is working as intended) and z_2 represents the operational standard that minimizes the number of decision errors, then the z-score corresponding to the judges' conceptual standard can be estimated by the formula[3]

$$z_y = \pm \sqrt{z_1 z_2}$$

where z_y takes the same sign as z_1 and z_2. To apply z_y, you need to know the distribution of the latent trait underlying the proficiency judgments. This variable is not observable, but you can reasonably assume that its distribution has the same shape as the distribution of the test scores.

To use this approach, it is necessary to know the operational standards x_1 and x_2 (both computed from the judgments of the same judges, if possible). One way to determine both standards in a single study is to conduct a person-judgment study in which the judges classify students into three categories: "meets the standard," "borderline," and "does not meet the standard." The judges should also have a fourth option, "cannot classify this student." (Otherwise, the judges may classify as borderline those students whose proficiency they cannot judge.) The resulting estimate of the percentage of students who meet the judges' conceptual standard should be substantially less biased than the estimate produced by applying either x_1 or x_2 alone.

---

[3] The derivation of this formula is shown in the Appendix.

BEST COPY AVAILABLE

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Appendix

DERIVATION OF FORMULAS

Let X represent the test score variable, and let Y represent the judges' concept of proficiency. Assume that the distributions of X and Y are known, that the regression of each variable on the other is linear, and that all conditional distributions of either variable, conditioning on the other, are symmetric. Let r represent the correlation of X and Y. Let y represent the judges' conceptual standard, and let p represent the proportion of the test takers whose proficiency is below it.

Let x_1 represent the operational standard implied by a borderline-group or item-judgment study. Then x_1 is the median of the conditional distribution of X, given $Y = y$. But if this conditional distribution is symmetric, its median is equal to its mean, which is given by the linear regression of X on Y, evaluated at $Y = y$. Therefore, if x_1 is the operational standard implied by a borderline-group or item-judgment study, then

$$x_1 = \mu_x + r\frac{\sigma_x}{\sigma_y}(y-\mu_y) \ . \tag{1}$$

If z_1 and z_y are the standardized scores corresponding to x_1 and y, then

$$z_1 = rz_y \ . \tag{2}$$

Let x_2 represent the operational standard that minimizes the number of decision errors. Then x_2 will be the test score at which half the test takers are above the judges' conceptual standard and half the test takers are below that standard. In other words, x_2 is chosen so that the conditional distribution of Y, given $X = x_2$, has the value y as its median. If this conditional distribution is symmetric, its median is equal to its mean. Its mean is given by the regression of Y on X, evaluated at $X = x_2$, and this conditional mean is equal to y, the judges' conceptual standard. Therefore,

$$\mu_y + r\frac{\sigma_y}{\sigma_x}(x_2-\mu_x) = y \ . \tag{3}$$

If z_2 and z_y are the standardized scores corresponding to x_2 and y, then

$$rz_2 = z_y \ . \tag{4}$$

You now have two equations that you can solve for r, to get

$$9r = \frac{z_1}{z_y} = \frac{z_y}{z_2} \ ; \tag{5}$$

$$z_y^2 = z_1 z_2 \ ; \tag{6}$$

$$z_y^2 = z_1 z_2 ; \tag{7}$$

where z_y takes the same sign as z_1 and z_2.

Standards for Reporting the Educational Achievement of Groups

Summary of Break-out Session[1]

The discussion focused on the *item-judgment and borderline-group approaches to standard setting* and on the different directions of bias inherent in the results of studies based on these approaches (as explained in Livingston's paper). The group examined ways to converge bias resulting from both methods by using the same conceptual and operational standards and also by the use of equating techniques. It was clarified that equating, when applied to methods of standard setting, requires the same set of judges to use both methods.

Livingston summarized the major sources of bias resulting from the two approaches to standard setting. He reviewed the theoretical and statistical approach to combining the operational standards, the results of which would be an estimate less biased than an estimate based on either of the two operational standards by itself. One participant suggested an alternative to combining the two approaches: Use one approach and identify those students whose test scores fall below the operational standard but whose actual proficiency is below the conceptual (i.e., true, as conceptualized by the standard setters) standard. This approach was evaluated as possible for a small group of test takers, since the conceptual standard is not directly measurable.

Participants posed some questions related to the two approaches. The first was: Can an operational (quantitative) standard be devised that would measure more precisely the conceptual (qualitative) standard? The second question related to the formula presented by Livingston where the z-scores corresponding to the two types of operational standards are represented by z_{x1} and z_{x2}:

$$z_y = \sqrt{z_{x1} z_{x2}} \qquad (8)$$

Is it possible to estimate both z-scores in the formula needed to estimate the conceptual standard, z_y? It was suggested that in order to narrow the distribution of test scores so that it conforms more accurately to the judges' concept of how the students are distributed, judges should be made to adjust their conceptual standard to what the test is measuring. Participants agreed that this approach was undesirable, because the judges might not agree on what is being measured. It was further suggested that conceptual standards could be defined at the time of test construction, rather than after the test had already been constructed. The judges could also be involved in generating or reviewing the test content specifications.

The discussion turned to *minimizing bias* when making decisions about individual students. It was suggested that the best approach was to minimize decision errors through lowering the operational standard. At the end of the discussion, the group concluded: (a) This approach should not be used to report estimated percentages of qualified students since it produces bias toward an overestimate of proficient students, and (b) a single method of setting standards cannot be optimal for both determining the percentage of a group that satisfies the standard and for making the correct decision for individual test takers.

---

[1] This is a summary of the break-out discussions for Samuel Livingston's presentation. The session was facilitated by Mary Crovo (National Assessment Governing Board) and recorded by Arnold Goldstein (National Center for Education Statistics).

The group explored bias that results from *imperfect test measurement of the judges' conceptual standard* in terms of the regression effect. The discussion focused on borderline students who are selected on the basis of the judges' conception, which in fact is not correlated exactly with actual test scores. Students below the conceptual standard mostly fall below the operational standard as measured by the test, but not as far below the operational standard relative to the group, because their scores regress to the mean.

This group emphasized that in order to arrive at a sound conceptual standard, *training of judges is crucial*. It was suggested that this training should include an explicit definition of the conceptual standard in terms that will allow the judges to translate it into an operational standard. Some participants submitted that one should strive for agreement on a conceptual standard by the judges; others indicated that judges disagree regardless of how hard one tries for agreement. Additionally, two procedures were suggested to establish agreement between judges: (a) Judges can share their concepts during the standard-setting process, and (b) they can be exposed to actual student performance. While there was no agreement as to the merit of either procedure, the group acknowledged their positive aspects, which were that sharing of conceptualizations might promote greater agreement and that exposure to student performance could serve to bring conceptual standards more into line with actual performance.

Participants expressed concern about *content coverage in the tests*, more specifically with the degree to which a test measures proficiency in all areas of the domain. It was noted that in contrasting-group studies, the assumption usually is that a test should classify a student's proficiency over the entire domain. With regard to *the composition of the tested population and the issue of subgroups*, it was suggested that the population of test takers can be viewed as a collection of subgroups at different levels along the test score distribution. While it was suggested that results for these subgroups could be reported separately, it was acknowledged that this would be more complicated and expensive than considering all test takers as a single population. In response to a question about *repeating the standard-setting process versus equating scores over time*, Livingston suggested that it is better to hold standards constant over a certain period of years to be able to derive year-to-year comparisons of results. Scores can be equated between tests by common-item or common-score methods. This would allow an updating of expectations of what students can do.

Participants from *state testing environments* presented their contexts of the standard-setting process. They pointed out that at the state level, standards are set in relation to academic performance, group data, and school accountability. Citing one case of Florida teacher certification, where there were only 10 test takers, they argued that the conditions surrounding the state tests might not satisfy the mathematical requirements of any method. The state participants further argued that there is not much guidance available regarding the political realities they must confront. Livingston responded that in cases like that of Florida, a person-judgment study should work if the tests are objective. He further recommended that at the state level, those responsible for standard setting should seek a wide variety of expert advice to support their effort.

The group then summarized points that should be addressed when constructing a defensible standard-setting process: (a) The selection process and qualification of judges should be carefully considered;

(b) judges should be fairly representative of the population of persons who are qualified to make judgments; (c) the standard-setting method should be one that the judges can use; (d) person-judgment methods work best because judges can compare examples of individual performance with their conceptual standard, and item-judgment methods are more difficult because the judges have no observations upon which to base their judgments; (e) given the different purposes of testing programs, it is important in some circumstances to measure year-to-year progress; and (f) cut scores are important in high-stakes testing.

On the Cognitive Construction of Standard-Setting Judgments: The Case of Configural Scoring

Richard M. Jaeger

Center for Educational Research and Evaluation
University of North Carolina, Greensboro

ABSTRACT

Setting performance standards can be regarded as an example of a class of problems identified in the psychological literature as judgment or decision-making (JDM) tasks. Such tasks are characterized by uncertainty of information or outcome or by an outcome dependent on personal preferences. Solutions to JDM tasks can be made more accurate by employing a variety of strategies that include increasing the clarity and completeness of problem statements, managed social interaction, and the use of decision aids. The relevant psychological literature is discussed, and an example of the use of such strategies in the context of a complex, multidimensional standard-setting problem is provided.

Setting performance standards is an example of a larger class of problems identified in the psychological literature as judgment or decision-making (JDM) tasks. A judgment or decision-making task is characterized by uncertainty of information or outcome, by an outcome dependent on personal preferences, or both.

JDM tasks are by nature complex. They evoke responses that are based on equally complex cognitive processes (Pitz & Sachs, 1984). Before responding to a JDM task, a person must encode the information that accompanies the task and develop a representation of the problem it presents. As Johnson-Laird (1981) described this process, the respondent develops a mental model that relates the problem to other knowledge.

In building a mental model, the respondent must address uncertainties generated by lack of completeness in the problem statement. These either must be resolved or represented in the respondent's mental model before a sound response can be formulated. In the case of performance standard setting, the uncertainties include, among others, (a) the ability distribution of examinees for whom a standard is to be set, (b) the difficulty of the test or assessment for which a standard is to be set, (c) the impact of the recommended standard on individuals and populations, (d) the implications of recommended standards for social institutions of which the respondent is a part, and (e) the steps to be followed in providing desired recommendations.

Tversky and Kahneman (1981), as well as Slovic, Fischoff, and Lichtenstein (1982), have shown that the wording of a JDM task has marked influence on the representation of uncertainties in the mental model a respondent creates and, in turn, on the judgment or decision a respondent provides. These grounded conclusions are consistent with many findings reported in the educational measurement literature on standard setting. It is well known that performance standards are substantially method dependent; indeed, in experiments involving several standard-setting methods, variance across methods sometimes exceeds respondent variance within method (see Jaeger, 1989, for a summary of this literature).

Responses to JDM tasks, including standard-setting tasks, are thus responses to problem statements that are replete with uncertainties and less-than-complete information. Berkeley and Humphreys (1982) suggested that a respondent's behavior in the face of such a task is a reflection of an attempt to resolve or cope with the uncertainties it presents. If in doubt about the applicability of this conclusion to performance-standard-setting problems, just recall the types of questions typically posed in the most frequently used standard-setting procedures. Examples include, How many of the response options to this item would a minimally qualified examinee reject as incorrect (Nedelsky, 1954)? What is the probability that a minimally qualified person would respond to this item correctly (Angoff, 1971)? Which students in your classes are right "at the border" that distinguishes competent from incompetent mathematics students (Livingston & Zieky, 1982)? It is very unlikely that respondents approach a standard-setting task prepared to deal with questions such as these.

How do respondents cope with the uncertainties JDM tasks present? According to Berkeley and Humphreys, convention respondents typically retrieve information from memory in an attempt to construct the missing links in their cognitive representation of the problem. Their mental models thus consist of explicit problem elements in addition to inferences that are based on information retrieved from general memory and cognitive anchors that locate the problem in a network of experience-based preconceptions. If the task presented is indefinite, the respondent must depend heavily on prior experience to formulate a response.

Uncertainty has been classified by Kahneman, Slovic, and Tversky (1982) as external or environmentally determined (for example, through the problem statement and its context) and internal or knowledge based (i.e., as outside or peripheral to the realm of past exposure, experience, or consideration). One way to reduce uncertainty is to present problems in terms that are linked to experience, either past or proximate. Griggs and Newstead (1982) have shown that a critical variable in the successful application of an abstract principle (such as a minimally qualified person) to the solution of a JDM task is whether the presentation of the problem leads to the construction of a mental model that is consistent with the principle. The less uncertainty created by the problem statement and its presentation, the more likely it is that consistency will be obtained.

One representation of the processes used by respondents to resolve uncertainties relies on the information-processing models of Anderson (1970) or Birnbaum (1982). Within this information-processing framework, the concept of anchoring and adjustment provides a useful conception of the sequence of mental operations respondents use to formulate a response to a JDM task. An anchor point that serves as an initial response to the problem is first constructed. This point is then adjusted in an amount and direction that is guided by the encoding of additional features of the problem and its presentation. A similar description of the processes used to solve predictive judgment problems was advanced by Rodrigo, de Vega, and Castaneda (1992). Pitz and Sachs (1984) suggested that this conception of the processes used to solve JDM tasks is very broadly applicable; citing Einhorn and Hogarth (1982), Lopes and Ekberg (1980), and Lopes (1982), they claim that it applies to any task in which a numerical response is required. It thus applies to standard-setting tasks.

A DECISION-MODEL APPROACH TO STANDARD SETTING

It would not be an exaggeration to characterize the body of research on setting performance standards as a repository of empirical studies on how standard-setting panelists' respond to alternative problem statements. Many studies have employed a single problem statement and a few have been comparative.

In the Campbell and Stanley (1963) lexicon, these studies mainly have been of the single-group "X O," treatment-followed-by-observation variety, and less often of the "X O" versus "X'O" variety that incorporates alternative treatments.

For the most part, if not exclusively, the experimental treatments (that is, the standard-setting methods) have been ad hoc constructions: seemingly good ideas proposed by their authors that are totally devoid of theoretical grounding. I would include in this blanket categorization my own offerings (e.g., Jaeger, 1982), as well as those of Nedelsky (1954), Angoff (1971) (including its many variants), and the model proposed by Ebel (1972). One could claim that the person-centered models of Livingston and Zieky (1982) are grounded in decision theory, but only loosely so. No real attention has been paid to the structure of loss functions, much less to the parameters of those functions.

Very few standard-setting panelists are likely to approach their tasks with a rich storehouse of prior experience or well-formulated conceptions of the decision alternatives standard setting presents. As noted earlier, uncertainties abound. To the extent that problem statements and standard-setting procedures leave respondents to their own devices, they will draw on their varied, and often limited, experiences and retrieve from memory information that is of inconsistent relevance to the problem they must solve. To the extent that this process of mental modeling can be guided through appropriate decision aids, results should be more consistent and more likely grounded in appropriate bases for resolution of uncertainty.

Appropriate resolution of uncertainty can be facilitated through sequential provision of information, followed by response testing. Guided social interaction can help panelists test their initial anchors and adjust them through comparison of the reasonableness of their justifications with those provided by others. Introduction of alternative cognitive schema can help panelists shape the mental models they adopt by making them aware of modes of organization that were outside their realm of prior experience. Standard setting can be conceptualized as a process that helps panelists to resolve uncertainties through the adoption of task-relevant strategies and information.

AN EXAMPLE

This paper was inspired by a standard-setting problem that is more complex than those advanced in the past in that it involves configural, rather than summative, scoring of relatively few performance exercises. Following is a description of one attempt to reduce uncertainties through structured provision of information, social moderation, and education that was designed to facilitate standard-setting-panelists' formulation of alternative mental models of the problem at hand and its solution.

The uncertainty-reduction strategies enumerated earlier were applied to the problem of establishing performance standards for exercises that compose an assessment of teachers seeking certification from the National Board for Professional Teaching Standards. Unlike traditional assessments, these exercises are polytomously scored and are not considered to be exchangeable components of a test that provides a unidimensional representation of some underlying ability. The performance of a candidate for certification is represented by a profile of scores, each on a structurally comparable scale, across the exercises that compose the assessment.

The standard-setting problem to be solved in this case requires specification of the subset of score profiles, selected from the domain of all possibilities, that should be associated with National Board

Certification of candidate teachers. In the case of one certification field, that of Early Adolescence Generalist teachers, the assessment was composed of seven exercises that were scored on a four-point scale. The number of potential score profiles was thus 4^7, or 16,384.

The procedure used to elicit standard-setting recommendations in this case was designed to facilitate respondents' construction of mental models of the problem and plausible alternative solutions. It incorporated clear definitions of decision alternatives, strategies to help panelists make manifest their initial anchors, guided social interaction, education on alternative conceptual models for structuring domains of acceptable score profiles, and feedback on the implications of interim recommendations.

The standard-setting panel was composed of 30 teachers who had completed all seven exercises of the Early Adolescence Generalist assessment. Panelists were selected on the basis of their superior performances on the assessment, with appropriate attention to regional, gender, ethnic, and racial diversity.

Two days were spent reminding panelists about the content and structure of the assessment exercises, teaching them about the procedures used to score each exercise, and teaching them the meaning of each score value on the four-point scale used to summarize candidates' performances. This instruction was provided by the developers of the exercises and scoring procedures: professional staff of the Performance Assessment Laboratory at the University of Georgia.

To reduce uncertainty associated with lack of knowledge about the difficulty of the test or assessment for which a standard is to be set, it is essential that standard-setting panelists become thoroughly familiar with the items or exercises that compose the assessment and the criteria that are to be used to judge the performances of examinees. In the case of a polytomously scored assessment, those who set performance standards must understand the meaning of each score value that could be assigned to examinees' performances. Hence the instructional procedures followed here.

Arguably, the best way to familiarize a standard-setting panel with the items or exercises that compose a test or assessment is to have them complete the assessment themselves. This would seem to be especially important when standards are to be set for complex performance assessments, since mere review of the exercises is unlikely to reveal their difficulties, and a high degree of uncertainty is likely to remain.

To help panelists to construct an anchor point for their mental models of appropriate standard-setting recommendations, they were introduced to the concepts of compensatory, conjunctive, and mixed standard-setting models. These models were proposed as alternative strategies for integrating profiles of performance on a set of performance exercises. Under a compensatory model, high levels of performance on some exercises can compensate for low levels of performance on others and still lead to superior evaluations of examinees' overall performances. Under a conjunctive model, high-level evaluations of examinees' overall performances can result only from a conjunction of high-level performances on all exercises. Finally, mixed models combine compensatory allowances for some subsets of exercises with conjunctive requirements for others.

Panelists applied these concepts by independently evaluating and classifying the profiles of performance of 200 hypothetical candidates for National Board Certification on seven performance exercises. They used a four-category scale, with anchor points labeled "superb," "accomplished," "competent," and "deficient," making use of previously provided definitions of these terms. Since the National Board for Professional Teaching Standards intends to certify "accomplished" teachers, the top two categories corresponded to success on the assessment. One of the 200 profiles of performance used in this procedure is shown in Figure 1.

Panelists' responses to the 200 profiles were used to complete a judgmental policy-capturing analysis. In this analysis, mathematical models were used in an attempt to make manifest the mental models panelists were applying to the standard-setting task. Variations of ordinary least squares regression were used to fit compensatory and conjunctive analytic models to the stimulus profiles and the response sets produced by each panelist. The compensatory model was of the form:

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \ldots + b_7X_7, \qquad (1)$$

where Ŷ denotes a predicted overall performance score, and X_i denotes a score on the ith exercise in a profile (i = 1, 2, ... , 7).

This model is termed compensatory because it is strictly additive. In this model an overall judgment depends on the (weighted) average quality of all attributes (Coombs, 1964). In the compensatory judgmental policy-capturing model, a high score on one exercise can compensate for a low score on another, to produce a predicted overall performance score at a given level. The conjunctive analytic model was of the form:

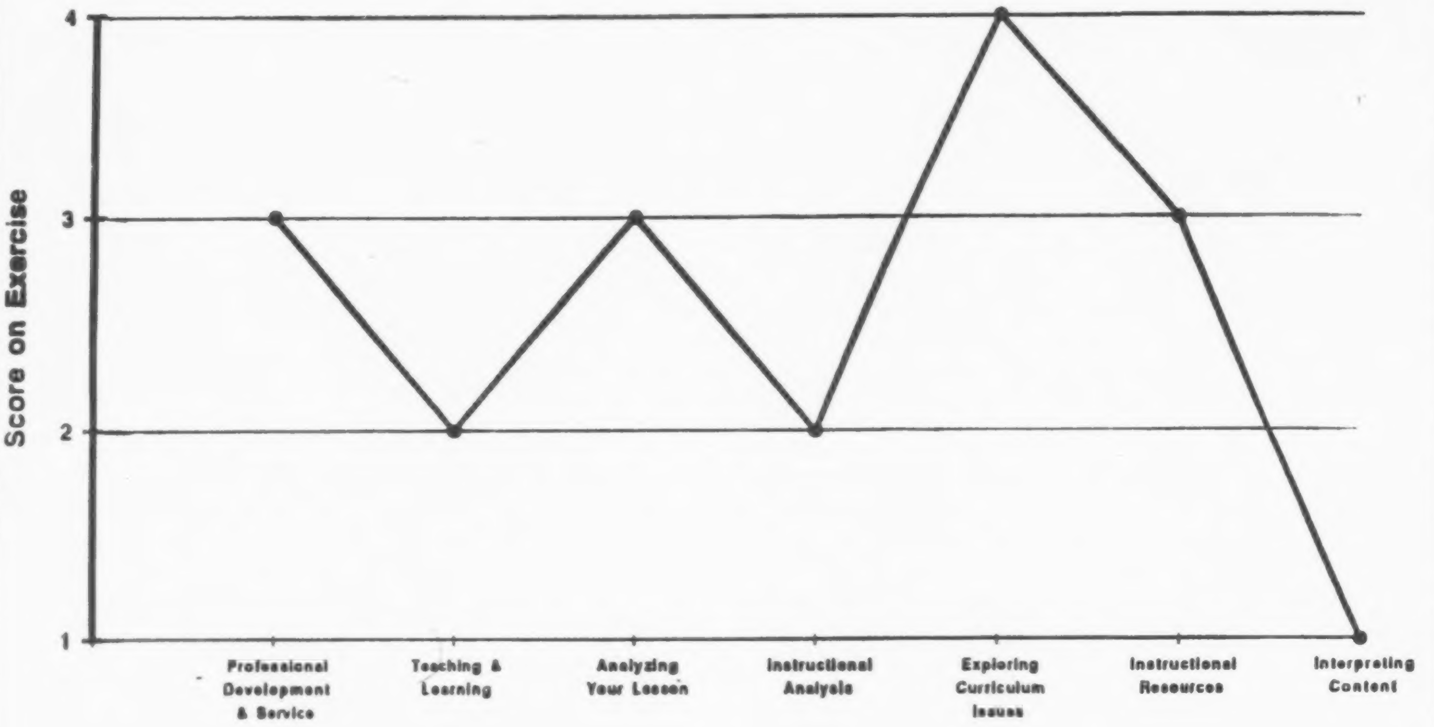
$$\hat{Y} = aX_1^{b_1}X_2^{b_2}\ldots X_7^{b_7} = a\prod_{i=1}^{7} X_i^{b_i}, \qquad (2)$$

where Ŷ denotes a predicted overall performance score, and X_i denotes a score on the ith exercise in a profile (i = 1, ... , 7).

The conjunctive model provides an approximation to a judgment policy in which high scores on all exercises are necessary to produce a high overall performance score. The parameters of a conjunctive model can be estimated using ordinary least squares multiple regression by applying a natural logarithmic transformation to the model prior to estimation. This produces a model of the form:

$$1n(\hat{Y}) = 1n(a) + \sum_{i=1}^{k} b_1 1n(X_i). \qquad (3)$$

To inform panelists about the characteristics of the anchor points that defined their initial mental models, they were provided with two forms of information. Instruction on the interpretation of the information also was provided. Based on the fitting of compensatory and conjunctive analytic models to their recommendations, panelists were given (a) the distribution of importance weights they attributed to each of the seven exercises that composed the assessment, (b) graphs showing how their distribution of weights compared to those of their fellow panelists, and (c) 100 profiles of performance on the seven exercises and the associated levels of overall candidate performance that were predicted by the analytic model. From these data, panelists could gain information about the structure of their mental models and learn about the severity of decisions that would result from application of those models. This kind of information is precisely the sort identified by Pitz and Sachs (1984) for reducing



Overall Evaluation of Candidate →

A ①	B ②	C ③	D ④
Deficient	**Competent**	**Accomplished**	**Superb**

Figure 1. Simulated profile of a candidate's performances on seven Early Adolescence Generalist assessment exercises.

Fill the appropriate bubble (A, B, C, or D) on the bubble sheet to indicate your judgment.
Be sure to match the response number on the bubble sheet to the profile number in the upper right hand corner (above).

the uncertainty of JDM tasks. An example of the kinds of information provided to standard-setting panelists is shown in Figure 2 and Table 1.

When used with care (Fitzpatrick, 1989), social interaction can help standard-setting panelists further reduce their uncertainties by determining whether their initial anchor points appear to be consistent with, or disparate from, those of others who purport to be equally qualified. Ashton (1992) found that having judges present justifications of their initial recommendations increased their judgment accuracy in situations where objective criteria were available and led to greater interjudge consensus in situations where they were not. For these reasons, a controlled, social-interaction procedure was used to elicit rationales from panelists and to inform them about the models held by their fellow panelists and the reasoning underlying those models. This procedure is consistent with the fundamental assumption that possession of relevant information that can be brought to bear on the decision process reduces uncertainty (Griggs & Newstead, 1982).

The effects of strategies to reduce uncertainty in judgment and decision processes can be realized only by providing subjects with opportunities to alter their initial judgments. For this reason, standard-setting panelists were allowed to replicate the judgmental, policy-capturing process, again categorizing the overall performances of 200 hypothetical candidates for National Board Certification, each represented by a seven-dimension profile of exercise scores.

As before, each panelist's responses were analyzed by fitting compensatory and conjunctive analytic models, and panelists were provided with information on the relative weights they and other panelists assigned to each of the seven exercises. They also were provided with tables containing profiles of performance on the seven exercises and the predicted overall performance category associated with each profile. Following discussion of this information, panelists proceeded to the second phase of the standard-setting process.

The grounded activities of the first phase of the standard-setting process--reduction of uncertainty through explicit problem definition, extensive training on the mechanics of decision aids, clarification of mental models through presentation of rationales, gaining information on the models held by others, and provision of information on the impact of intermediate recommendations--were carried over to the second phase. The difference between the two phases was the focus on elicitation of implicit mental models during the first phase, and the focus on development of explicit model statements during the second phase.

Standard-setting panelists were provided with a detailed reminder of the National Board for Professional Teaching Standards' definition of an accomplished teacher who was worthy of certification. They were then asked to specify the characteristics of performances on the seven exercises of the assessment that were just above the threshold that separated teachers who satisfied the National Board's definition and those who did not. Panelists were invited to respond to this task by writing narrative statements and by providing samples of performance profiles that they judged to be just marginally consistent with the National Board's definition of accomplished Early Adolescence Generalist teaching.

Two iterations of this process were separated by structured discussions in which panelists described their explicit models of accomplished performance on the seven assessment exercises and provided justifications for their recommendations. Panelists thus made their models known to others and learned about the models an,d justifications voiced by their colleagues. Following the second iteration, the

PERCENT OF VARIANCE EXPLAINED = 53%

THE RELATIVE WEIGHTS YOU ASSIGNED TO EXERCISES

0.21	Professional Development and Service (PDS)
0.18	Teaching and Learning (T&L)
0.14	Analyzing Your Lesson (AYL)
0.18	Instructional Analysis (IA)
0.08	Exploring Curriculum Issues (ECI)
0.06	Instructional Resources (IR)
0.14	Interpreting Content (IC)

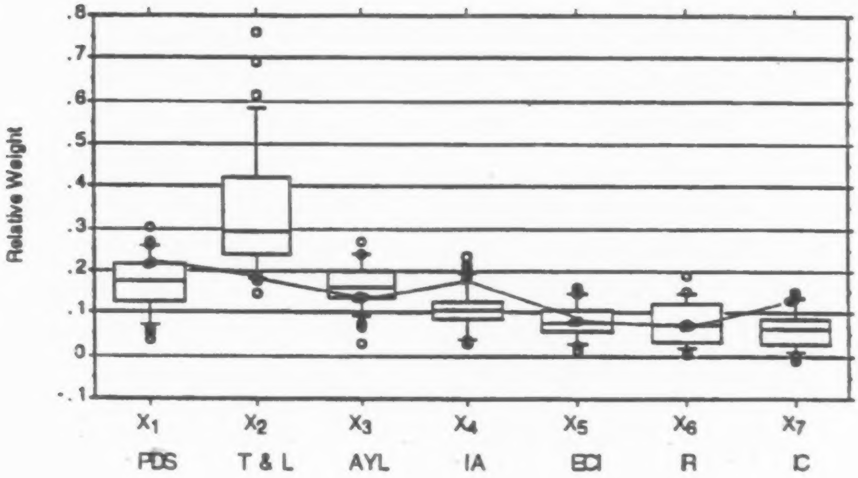


Figure 2. Distributions of relative weights assigned to the seven Early Adolescence Generalist exercises by the standard-setting panel of 300 teachers.

Table I

A Sample of Candidate Classifications and Certification Decisions Resulting From the Standard-Setting Policy Proposed by a Single Member of the Standard-Setting Panel

	SCORE LEVEL ON EACH EXERCISE								
PROFILE NUMBER	PROFESSIONAL DEVELOPMENT AND SERVICE	TEACHING AND LEARNING	ANALYZING YOUR LESSON	INSTRUCTIONAL ANALYSIS	EXPLORING CURRICULUM ISSUES	INSTRUCTIONAL RESOURCES	INTERPRETING CONTENT	OVERALL "SCORE" DERIVED BY APPLYING YOUR POLICY	DECISION ACCORDING TO YOUR POLICY (CERTIFY/NOT CERTIFY
1	4	4	4	4	4	4	4	4	CERTIFY
2	4	4	4	4	4	3	4	4	CERTIFY
3	4	3	4	4	3	4	4	4	CERTIFY
4	4	3	4	3	4	4	3	4	CERTIFY
5	3	3	4	3	4	4	4	3	CERTIFY
6	4	2	3	4	4	4	4	3	CERTIFY
7	2	3	4	4	4	4	4	3	CERTIFY
8	3	4	4	4	3	3	3	4	CERTIFY
9	3	3	3	4	3	4	4	3	CERTIFY
10	4	4	2	4	4	4	2	3	CERTIFY
11	4	3	2	4	4	3	4	3	CERTIFY
12	2	2	4	4	4	4	4	3	CERTIFY
13	3	4	4	2	4	3	3	3	CERTIFY
14	3	2	3	4	4	4	3	3	CERTIFY
15	2	4	3	4	4	2	4	3	CERTIFY
16	4	3	4	3	4	1	4	3	CERTIFY
17	4	3	3	3	3	3	3	3	CERTIFY
18	3	4	3	3	3	3	3	3	CERTIFY
19	2	4	3	4	3	2	4	3	CERTIFY
20	4	4	3	3	1	4	3	3	CERTIFY
21	3	1	4	4	2	4	4	3	CERTIFY
22	3	4	4	3	2	2	3	3	CERTIFY
23	3	4	2	3	2	3	4	3	CERTIFY
24	3	2	3	3	4	3	3	3	CERTIFY
25	3	3	4	1	3	3	4	3	CERTIFY
26	3	1	2	4	4	4	3	3	CERTIFY
27	3	3	3	3	3	3	3	3	CERTIFY
28	4	4	3	2	3	2	2	3	CERTIFY
29	4	2	4	3	2	3	2	3	CERTIFY
30	4	2	2	3	3	3	3	3	CERTIFY
31	3	2	4	2	3	2	4	3	CERTIFY
32	4	4	3	1	3	4	1	2	DON'T CERTIFY
33	4	4	1	4	2	2	3	3	CERTIFY
34	1	2	2	4	3	4	4	3	CERTIFY
35	3	2	3	3	3	3	3	3	CERTIFY
36	2	3	3	3	3	3	3	3	CERTIFY
37	4	2	2	3	3	3	2	3	CERTIFY
38	3	4	3	2	2	2	3	3	CERTIFY

coprincipal investigators of the study completed a content analysis of the panelists' recommended performance standards and identified four compensatory and conjunctive elements that were common to almost all of them. Four standards emerged from this analysis.

To help panelists understand the impact of applying each of these summary performance standards, they were provided with information on the percentage of candidates for National Board Certification whose profiles of performance on the seven exercises would satisfy each recommended standard. This information was provided to address a significant source of residual uncertainty that might cause panelists to recommend standards that they would later consider to be unreasonable.

With this source of uncertainty resolved, panelists were asked to discuss the information provided to them and then to reconsider the standards they had previously recommended. Following their discussion, panelists constructed modified performance standards by again specifying levels of performance on the seven exercises of the assessment that were, in their judgment, just above the threshold that separated teachers who satisfied the National Board for Professional Teaching Standards' definition of accomplished teaching and those who did not.

The coprincipal investigators of this study again completed a content analysis of panelists' recommended performance standards. They found that four alternative standards captured the essential features of the recommendations provided by almost all of the 30 panelists.

At this point, panelists were invited to discuss these four alternatives and to propose modifications that were more consistent with their conceptions of standards that distinguished accomplished from less-than-accomplished Early Adolescence Generalist teachers. Suggested modifications were discussed by the entire panel and the acceptability of these modifications was determined through formal voting by panel members.

When all recommendations had been adopted or rejected, panelists were asked to rank order the four modified performance standards in terms of preference and to express their views concerning the appropriateness of the certification decisions that would result if each standard were adopted by the National Board for Professional Teaching Standards. When these results were presented to the panel, a motion to adopt the most popular of the four recommended standards was introduced by one panelist. In a subsequent vote, the motion was supported by 87% of the panel and opposed by only 3%. Ten percent of panelists expressed neither support nor opposition.

SUMMARY AND CONCLUSION

The educational measurement literature on performance standard setting consists primarily of proposals of alternative stimulus statements, reports on the application of those statements, and comparisons of the results of applying two or more such statements. With rare and arguable exception, these proposals are ad hoc ideas presented by their authors in the absence of a conceptual framework and are devoid of reference to the psychological literature on judgment and decision processes. (See Berk, 1986; Jaeger, 1989; and Hambleton & Eignor, 1980, for reviews and summaries of this literature.)

This paper proposes a reconceptualization of standard setting through the recognition that standard-setting tasks are examples of a larger category of judgment and decision problems. These problems have been addressed in a rich and varied body of psychological research that has led to the formulation of models of the cognitive processes subjects use to address the uncertainties all such tasks present. By taking these cognitive processes into account when developing and using standard-setting procedures, the uncertainties surrounding all standard-setting tasks can be reduced in ways that will facilitate the development of reasoned standard-setting recommendations.

Strategies that are consistent with the psychological literature were illustrated in the context of a standard-setting problem that required consideration of profiles of performance on a set of performance assessment exercises. The multiphase procedure used in this example facilitated panelists' formulation of initial mental models of the standard-setting problem and its solution. The procedure then guided the adjustment of panelists' models through a variety of devices that helped them reduce residual uncertainties in reasoned, factually grounded ways.

References

Anderson, N. H. (1970). Functional measurement and psychophysical judgment. *Psychological Review, 77,* 153-170.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Ashton, R. H. (1992). Effects of justification as a mechanical aid on judgment performance. *Organizational Behavior and Human Decision Processes, 52,* 292-306.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56,* 137-172.

Berkeley, D., & Humphreys, P. (1982). Structuring decision problems and the bias heuristic. *Acta Psycholgica, 50,* 201-252.

Birnbaum, M. H. (1982). Controversies in psychological measurement. In M. H. Birnbaum (Ed.) *Social attitudes and psychological measurement* (pp. 401-485). Hillsdale, NJ: Erlbaum.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research in teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally.

Coombs, C. (1964). *A theory of data*. New York: Wiley.

Ebel, R. L. (1972). *Essentials of educational measurement* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Einhorn, H. J., & Hogarth, R. M. (1982). *A theory of diagnostic inference II.* Chicago: University of Chicago, Graduate School of Business, Center for Decision Research.

Fitzpatrick, A. R. (1989). Social influences in standard setting: The effects of social interaction on group judgments. *Review of Educational Research, 59,* 315-328.

Griggs, R. A., & Newstead, S. E. (1982). The role of problem structure in a deductive reasoning task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 297-307.

Hambleton, R. K., & Eignor, D. R. (1980). Competency test development, validation, and standard setting. In R. M. Jaeger & C. K. Tittle (Eds.), *Minimum competency achievement testing: Motives, models, measures, and consequences* (pp. 367-396). Berkeley, CA: McCutchan.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4*(4), 461-475.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). New York: American Council on Education/Macmillan.

Johnson-Laird, P. N. (1981). Mental models in cognitive science. In D. A. Norman (Ed.), *Perspectives on cognitive science* (pp. 147-191). Hillsdale, NJ: Erlbaum.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Lopes, L. L. (1982). *Toward a procedural theory of judgment*. (Tech. Rep. No. 17). Madison: Wisconsin Human Information Processing Program.

Lopes, L. L., & Ekberg, P. H. S. (1980). Test of an ordering hypothesis in risky decision making. *Acta Psychological, 45,* 161-167.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Pitz, G. F., & Sachs, N. J. (1984). Judgment and decision: Theory and application. *Annual Review of Psychology, 35,* 139-163.

Rodrigo, M. J., de Vega, M., & Castaneda, J. (1992). Updating mental models in predictive reasoning. *European Journal of Cognitive Psychology, 4,* 141-157.

Slovic, P., Fischoff, B., & Lichtenstein, S. (1982). Response mode, framing, and information-processing effects in risk assessment. In R. Hogarth (Ed.), *New directions for methodology of social and behavioral science: Question framing and response consistency* (pp. 21-36). San Francisco: Jossey-Bass.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453-458.

On the Cognitive Construction of Standard-Setting Judgments: The Case of Configural Scoring

Summary of Break-out Session[1]

The discussion centered around three broad areas: (a) criteria for standards and the process of setting standards, (b) judges' behavior and mathematical modeling of decision making, and (c) standards and the communication of assessment results to the public. The group affirmed the major desirable characteristics of standards: appropriateness, internal consistency, and reproducibility both over judging occasions and among judges at a single point in time. In a discussion of consistency and reproducibility, the participants noted that although it is possible to remove inconsistent judges from a panel, removal rarely occurs. Participants also explored how conceptual standards are determined. The group concluded that this process is usually based on the goal and purpose of the test as well as on the distribution of abilities among the test takers.

The discussion then focused on the *methods for setting standards*: item options (Nedelsky), probabilities (Angoff), and contrasting-group studies. Participants sought to clarify when either the item or the whole-booklet methods should be used. It was indicated that different methods lead to different standards.

One participant questioned whether there were approaches other than having panels of judges set standards. This participant was interested in exploring predictive methods that use a desired level of occupational performance as a standard to determine what knowledge or skills are needed to attain that level. It was suggested that if accepted external criteria exist (e.g., performance level is known), then previously administered assessments could be used to see what items predict success in the real world. It was emphasized that this is a process for the selection of students, not for setting standards.

Two questions guided the discussion related to the *process of implementing a program of standards*: (a) Which comes first, setting standards or the assessment? and (b) Can we build a model of characteristics of assessment without first having a model for standard setting? With respect to the first question, there were multiple suggestions. Some suggested that the assessment should be done first. Others indicated that it did not matter, while others indicated that the process should be iterative. The essential element was that there ought to be a good match between them. In exploring the iterative process, participants expressed concern about a small number of items at the highest levels in the National Assessment of Educational Progress (NAEP), where a paucity of information may lead to inaccuracies. It was recommended that in an iterative process, one should start with ascertaining where standards should be set, use that information as a basis for designing the assessment, then devise questions near the cut-off point(s), so that a lot of information is centered around the level of the standards.

---

[1] This is a summary of the break-out discussions for Richard Jaeger's presentation. The session was facilitated by Andrew Kolstad (National Center for Education Statistics) and recorded by Mary Naifeh (National Center for Education Statistics).

The group focused briefly on three related areas: first, the need in standards setting and in assessment to ascertain not only what students know, but also to learn what cognitive processes they use to solve problems; second, the qualification of judges relative to the content area being assessed and the population from which they are polled; and third, the issue of setting standards for ethnic and language minority students. The participants agreed that standard-setting panels engage in complex cognitive processes involving judgment and decision making. They also reviewed and attested to the difficulty standards setters have in understanding the models used to make judgments. They agreed that to resolve uncertainties, panelists must be trained in (a) task-relevant strategies and information and (b) the nature and use of mathematical models.

In an extension of the discussion of mathematical models, the participants concurred that one challenge is to identify additional mathematical models of judgment making. There were difficulties inherent in the use of these models. For example, some models match decision making very well, while others do not. Poor matches may occur for many reasons. For example, the panelists are still developing the process of decision making and have not converged on a model, or the model may be incorrect.

The participants discussed issues of *reliability* as related to standard setting, focusing on the *consistency of a panelist's rating over a large number of examinees*. The difficulty in this situation relates to distinguishing whether the standard-setting process uses a very complex model that is not defined well, or whether the process is random. One participant indicated that in research on ability tests during the 1970s, consistent judgments by a panelist over time, coupled with a poor fit of the model, showed that the panelist's replicable behavior was not captured by the model. It was emphasized, though, that there exists a class of models for which sources of difficulty are evident. However, one participant suggested that some panelists never become consistent and, therefore, recommended that rules for eliminating inconsistent panelists as judges are needed. The discussion extended to the various elements of inconsistency: (a) The math model does not fit the judge in question; (b) the panelist is unable to articulate reasons for ostensibly inconsistent scores simply because the language is not sufficiently complex.

The group also focused on a second reliability issue, consistency over time. It was noted that neither this issue nor variations among judges' ratings has been sufficiently evaluated. It was pointed out that research on giving instructions to panelists shows: (a) That if explicit instructions are given, there is less inconsistency; and (b) different methods for setting standards over time will result in different standards being defined, even though panelists are not changed.

The following broad issues were also discussed for short periods: the role of memory and tradition in setting standards, the underlying assumption of rationality, and the lack of fit between dichotomous responses and the mathematical model.

The group then focused its discussion on *reporting NAEP results* to the public. The group, while commending the positive effects of reporting, expressed concern about the interpretations given to the public by the press. They suggested that part of the problem occurs because education editors and reporters are not well informed themselves about testing and measurement issues. Additionally, it was suggested that the public possesses minimum knowledge on the issues. In addition, the group criticized the NAEP's standards and reporting systems. It was suggested that Linn's research comparing the old anchor levels with today's standards shows little difference between the two; that is, the new standards report nothing different than the old anchor levels did, and the labels--Basic, Proficient, and Advanced--may mislead the public. It was argued that NAEP's original purpose and value was time series; the change to standards, where time series take on a different meaning, poses a tremendous risk of changing motivations for states. The group concluded that it is necessary to educate the public to understand NAEP results.

Some Technical Aspects of Standard Setting

Huynh Huynh[1]

Professor, College of Education, University of South Carolina

ABSTRACT

This paper deals with some technical aspects in standard setting. First, statistical accuracy and minimum sample size are considered for cut scores based on the contrasting-groups procedure. Next, a number of psychometric procedures are presented for mapping test score levels of performance assessment items, along with implications for standard-setting methods based on item judgment. This is followed by a decision-theoretic approach to adjusting cut scores for situations where adverse consequences of incorrect decisions need to be factored. Finally, remarks are made on practical issues such as local item dependence, number of score levels, and stability of standards in yearly equatings.

INTRODUCTION

As noted by many authors, including Jaeger (1989) and Livingston and Zieky (1982), most standard-setting methods in education can be categorized as either examinee-centered or test-centered. The most commonly used examinee-centered procedure is based on contrasting groups. A typical use of this model involves an external (i.e., independent of the test data) sorting of examinees into two or more groups. The cut score is then set at a point that is most consistent with the examinees' classifications. In a typical test-centered model such as the Angoff procedure, judges are asked to determine the level of performance on the test items or tasks that they consider as *barely adequate*. A different version of the test-centered model is used in the Maryland School Performance Assessment Program (MSPAP). In this case, judges are asked to identify test-based activities that can or cannot be accomplished by examinees in each predefined level of performance (MSPAP, 1994).

This paper focuses on some statistical aspects associated with both the examinee-centered and the test-centered models for standard setting. First, statistical accuracy and minimum sample size are considered for cut scores based on the contrasting-groups procedure. Next, a number of psychometric procedures are presented for mapping test score levels of performance-assessment items, along with implications for standard-setting methods based on item judgment. This is followed by a decision-theoretic approach to the adjustment of cut scores for situations where adverse consequences of incorrect decisions need to be factored. Finally, remarks are made on practical issues such as local item dependence, number of score levels, and stability of standards in yearly equatings.

---

[1] Many issues discussed in this paper originated from the author's work with the South Carolina Basic Skills Assessment Program and the Maryland School Performance Assessment Program. Gratitude is extended to Vana Meredith-Dabney, Steve Ferrara, Lynn Mappus, Paul Sandifer, and Joe Saunders for their many beneficial discussions over the years.

BEST COPY AVAILABLE

SAMPLE SIZE ISSUES FOR THE CONTRASTING-GROUPS PROCEDURE

The Normal Model

Assume that the test score X follows a normal distribution with mean μ_1 and standard deviation σ_1 for the lower or nonadequate group (Group 1) and a normal distribution with mean μ_2 and standard deviation σ_2 for the upper or adequate group (Group 2). The probability density functions (pdf) for these two groups may be written as

$$f_1(x) = (1/\sigma_1\sqrt{2\pi})\ exp\ [-(x - \mu_1)^2/2\sigma_1^2] \qquad (1)$$

and

$$f_2(x) = (1/\sigma_2\sqrt{2\pi})\ exp\ [-(x - \mu_2)^2/2\sigma_2^2]. \qquad (2)$$

Let c be a cut score. Then the probability of a false positive error is $P_1\ (X \geq c)$ computed under the pdf $f_1(x)$, and the probability of a false negative error is $P_2\ (X < c)$ computed under the pdf $f_2(x)$.

Denote Q as the ratio reflecting the importance of passing an adequate examinee to the importance of failing a nonadequate examinee. For example, if passing an adequate student is less important than failing a nonadequate student, then Q should be set at less than 1. Note that Q is also the ratio of the "loss" incurred by a false-negative error to the loss incurred by a false-positive error.

Assume that there are N_1 nonadequate students and N_2 adequate students and that the aim is to set a cut score that will minimize the expected loss for the combined $(N_1 + N_2)$ students. The expected loss in this case is $N_1P_1\ (X \geq c) + QN_2P_2\ (X < c)$. The cut score c at which the expected loss is at its extremum (maximum or minimum) satisfies the equality $N_1f_1(c) = QN_2f_2(c)$ or $f_1(c) = Q^*f_2(c)$, where $Q^* = QN_2/N_1$. By taking the logarithm of both sides of this equation, it may be verified that the cut score c is a solution of the equation

$$[(c - \mu_1)/\sigma_1]^2 - [(c - \mu_2)/\sigma_2]^2 + ln\sigma_1^2 - ln\sigma_2^2 + lnQ^* = 0. \qquad (3)$$

Now, let

$$\alpha = 1/\sigma_1^2 - 1/\sigma_2^2, \qquad (4)$$

$$\beta = \mu_1/\sigma_1^2 - \mu_2/\sigma_2^2, \text{ and} \qquad (5)$$

$$\gamma = \mu_1^2/\sigma_1^2 - \mu_2/\sigma_2^2 + ln(Q^*\sigma_1^2/\sigma_2^2). \qquad (6)$$

In the above expression, the notation *ln* represents the natural logarithm function. If the two standard deviations σ_1 and σ_2 are equal to the same value σ, then $\alpha = 0$ and the cut score is $c = [(\mu_1 + \mu_2) + \sigma^2 ln\ Q^* / (\mu_1 - \mu_2)]/2$. Otherwise, the cut score c is a solution of the quadratic equation

$$\alpha c^2 - 2\beta c + \gamma = 0. \tag{7}$$

Asymptotic Sampling

To study the asymptotic sampling distribution of the cut score c when sample data are used, consider the following function of c, μ_1, μ_2, σ_1^2, and σ_2^2:

$$F = [(c - \mu_1)/\sigma_1]^2 - [(c - \mu_2)/\sigma_2]^2 + ln\sigma_1^2 - ln\sigma_2^2 + lnQ^* = 0. \tag{8}$$

Then the cut score c is a solution of the equation $F(c) = 0$. Now use the notation $F'(u)$ to denote the partial derivative of the function F with respect to the variable u, where u may be c, μ_1, μ_2, σ_1^2, or σ_2^2. In addition, the cut score c can be written as a function of these means and standard deviations:

$$c = \varphi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2). \tag{9}$$

Now let the two standard scores z_1 and z_2 be defined as

$$z_1 = (c - \mu_1)/\sigma_1 \text{ and } z_2 = (c - \mu_2)/\sigma_2. \tag{10}$$

Then it may be verified that the partial derivatives of φ with respect to the means and variances are given as

$$\delta\varphi/\delta\mu_1 = -F'(\mu_1)/F'(c) = -2z_1/\sigma_1 F'(c) \tag{11}$$

$$\delta\varphi/\delta\mu_2 = -F'(\mu_2)/F'(c) = 2z_2/\sigma_2 F'(c) \tag{12}$$

$$\delta\varphi/\delta\sigma_1^2 = -F'(\sigma_1^2)/F'(c) = (-z_1^2 + 1)/\sigma_1^2 F'(c) \tag{13}$$

$$\delta\varphi/\delta\sigma_2^2 = -F'(\sigma_2^2)/F'(c) = (z_2^2 - 1)/\sigma_2^2 F'(c). \tag{14}$$

In the above expression,

$$F'(c) = 2(z_1/\sigma_1 - z_2/\sigma_2). \tag{15}$$

Now consider the case where the means and variance parameters of the test score of the lower and upper populations are estimated by the traditional sample means $\overline{x_1}$, $\overline{x_2}$ and sample (unbiased) variances s_1^2 and s_2^2 of a random sample of n_1 subjects from the lower group and a random sample of n_2 subjects from the upper group. Under the normality assumption, these estimates are stochastically independent and have sampling variances of

$$Var(\bar{x}_1) = \sigma_1^2/n_1 \tag{16}$$

$$Var(\bar{x}_2) = \sigma_2^2/n_2 \tag{17}$$

$$Var(s_1^2) = \sigma_1^4/(n_1-1) \tag{18}$$

$$Var(s_2^2) = \sigma_2^4/(n_2-1). \tag{19}$$

It follows that the estimated cut score $\hat{c}$ based on the sample data is asymptotically unbiased and has an asymptotic variance of

$$\begin{aligned} Var(\hat{c}) = {} & (\delta\varphi/\delta\mu_1)^2\ Var(\bar{x}_1) + (\delta\varphi/\delta\mu_2)^2\ Var(\bar{x}_2) \\ & + (\delta\varphi/\delta\sigma_1^2)^2\ Var(s_1^2) + (\delta\varphi/\delta\sigma_2^2)^2\ Var(s_2^2). \end{aligned} \tag{20}$$

Algebraic manipulations will yield

$$Var(\hat{c}) = 2[2z_1^2/n_1 + 2z_2^2/n_2 + (z_1^2 - 1)^2/(n_1 - 1) + (z_2^2 - 1)^2/(n_2 - 1)]/[F'(c)]^2, \tag{21}$$

where

$$F'(c) = 2(z_1/\sigma_1 - z_2/\sigma_2). \tag{22}$$

Note that, for c to be the cut score based on the contrasting-groups procedure, the partially derivative $F'(c)$ must be negative at the cut score c.

Of course, only sample data are available to the standard-setting process. Therefore, as in most estimation situations, the mean and standard deviation parameters may be replaced in the formula for the sampling variance of the estimated cut score $\hat{c}$ by the corresponding sample mean and sample variance.

Sample Size Consideration

To simplify the study of minimum sample size in the contrasting-groups procedure, consider only the case where the number N_1 of nonadequate students and the number N_2 of adequate students are equal and where the loss ratio Q is equal to 1. This situation is analogous to the statistical decision problem involving only one examinee for whom the pdf f_1 and f_2 serve as prior pdf and the errors of misclassification are equally serious.

For the above case, assume that the two samples have the same size ($n_1 = n_2 = n$). By taking n-1 as roughly equal to n, the sampling variance of the estimated cut score $\hat{c}$ can be written as

$$Var\ (\hat{c}) = G/n, \tag{23}$$

where G is free of n and is equal to

$$G = [z_1^2 + z_2^2 + (z_1^2 - 1)^2/2 + (z_2^2 - 1)^2/2]/(z_1/\sigma_1 - z_2/\sigma_2)^2. \tag{24}$$

Therefore, something must be known about the value of G and the magnitude of the sampling variance for $\hat{c}$ before a minimum sample size n for both the lower and upper groups can be determined.

For example, it may be reasonable for the case on integer test scores to require that the 95% band for the confidence interval for $\hat{c}$ not exceed 1. Under the normal asymptotic theory, n must satisfy the inequality

$$(2 \times 1.96)^2\ (G/n) < 1$$

or that

$$n > 15.4G. \tag{25}$$

Some Benchmarks

To appreciate how much the constant G might be in a large-scale assessment, consider the data collected in the May 1981 statewide administration of the tests of reading and mathematics of the South Carolina Basic Skills Assessment Program (BSAP) (Huynh & Casteel, 1983). Prior to the 1981 testing, samples of intact classes (the teacher-judgment samples) were selected and teachers were asked to evaluate the adequacy of their students' performance and to sort the students into three groups labeled Nonadequate (lower group), Adequate (upper group), or Undecided. Some summary data for students in the teacher-judgment samples are reported in Tables 1 and 2. These tables also show the minimum sample size n. It appears from the above derivations that the minimum sample size n varies with the variance s^2 of the combined group and with the standardized distance $d = (\bar{x}_2 - \bar{x}_1)/s$. For the 10 BSAP cases, the minimum sample size correlates negatively with d ($r = -.36$), positively with s^2 ($r = .77$), and relates to both d and s^2 with a multiple correlation of $R^2 = .92$.

The regression equation is

$$n = 480 - 347d + 4.8s^2, \tag{26}$$

and the standard error is 32. Larger samples are needed when there is more variability in the test data and when the means of the lower and upper groups are closer to each other. Conversely, smaller samples are sufficient when the two group means differ more from each other and when the test scores are more homogeneous. For situations similar to the 1981 BSAP data, it appears that the required sample size ranges from about 100 to as much as 400.

A Few Remarks

Note that the contrasting-groups procedure actually is a special case of a general model proposed by Huynh (1976) for the evaluation of true mastery scores when a referral task is available. This may be

Table 1

Descriptive Statistics for Contrasting-Groups Samples and Minimum Sample Size (MSS), BSAP Reading Test

Grade	Content	MSS	Statistics	Combined sample*	Non-adequate group	Adequate group
1	Reading	274	N	2,923	892	1,779
			Mean	26.08	19.13	29.99
			SD	7.76	5.95	5.91
2	Reading	304	N	2,675	862	1,636
			Mean	26.80	19.83	30.68
			SD	7.92	6.82	5.62
3	Reading	273	N	2,725	1,025	1,537
			Mean	27.57	22.42	31.24
			SD	7.17	6.99	4.66
6	Reading	283	N	2,677	1,012	1,422
			Mean	24.36	18.50	28.54
			SD	7.56	6.27	5.58
8	Reading	387	N	2,624	824	1,626
			Mean	24.40	17.99	27.68
			SD	7.84	6.80	6.19

Note. From *Technical Works for Basic Skills Assessment Programs* (Final Report of Grant NIE-G-90-0129), by H. Huynh and J. Casteel, 1983, p. 7.

*Including students in the Undecided group and others without grouping data.

Table 2

Descriptive Statistics for Contrasting-Groups Samples and Minimum Sample Size (MSS), BSAP Math Test

Grade	Content	MSS	Statistics	Combined sample*	Non-adequate group	Adequate group
1	Math	105	N	2,923	589	2,125
			Mean	25.32	20.69	26.74
			SD	4.16	4.61	2.87
2	Math	123	N	2,672	629	1,866
			Mean	25.87	22.71	27.00
			SD	3.72	4.25	2.85
3	Math	194	N	2,722	838	1,714
			Mean	22.65	19.29	24.38
			SD	4.70	4.42	3.87
6	Math	220	N	2,681	1,057	1,437
			Mean	16.16	12.21	19.97
			SD	6.34	4.68	5.42
8	Math	414	N	2,631	1,040	1,418
			Mean	13.41	10.11	15.73
			SD	6.48	4.74	6.55

Note. From *Technical Works for Basic Skills Assessment Programs* (Final Report of Grant NIE-G-90-0129), by H. Huynh and J. Casteel, 1983, p. 8.

*Including students in the Undecided group and others without grouping data.

seen by letting the loss ratio $Q = 1$ and the referral success be the conditional probability

$$s = f_2/(f_1 + f_2).$$

The analysis carried out here for cut scores based on the contrasting-groups procedure makes use of the assumption of normal test score distributions. Most National Assessment of Educational Progress (NAEP) data appear to approximate this requirement. In many situations, however, a statistical process is typically used to smooth out the pdfs for the lower and upper groups without imposing a parametric model (such as the normal distribution), and the cut score is found by letting the two pdfs intersect. Generally these situations require that each pdf be expressed as a function of more than two parameters (such as mean and standard deviation). Hence, it seems reasonable to assume that the resulting cut score would show more sampling variability than is predicted under the normal theory. As a consequence, these cases would require samples larger than those stipulated for normal distributions.

LOCATING A TWO-PARAMETER LOGISTIC (2PL) BINARY ITEM AND ITS 0 AND 1 SCORES

General Considerations

Consider now a binary item with the probabilities for the correct response ($X = 1$) and incorrect response ($X = 0$) governed by the (Birnbaum, 1968) 2PL model. With X as the test score, use the notations

$$P(\theta) = P(X = 1) = exp\ a(\theta - b)/[1 + exp\ a(\theta - b)] \text{ and} \tag{27}$$

$$Q(\theta) = P(X = 0) = 1/[1 + exp\ a(\theta - b)]. \tag{28}$$

The item parameter b is known to be the item location and is the place on the latent trait where the probabilities P and Q are equal. In some way the item location b represents the boundary between the scores $X = 0$ and $X = 1$ on the latent trait. Recall also that the item information is

$$I(\theta) = P(\theta)Q(\theta) = P(1 - \mathrm{P}) \text{ and is maximized at } \theta = b, \text{ where } P = Q = 1/2.$$

Now, the information for the score level $X = 0$ is equal to $I_0(\theta) = PQ^2$ and is maximized at θ, at which $Q(\theta) = 2/3$ (= 67%). It follows that the maximum information (MI) score level location of $X = 0$ is equal to $\tau_0 = \mathrm{b} - a/ln\ 2$, where $ln\ 2$, is the natural logarithm of 2 (or approximately .693). As for the score level $X = 1$, the information is $I_1(\theta) = P^2Q$ and is maximized when $P(\theta) = 2/3$. Thus, the MI score level location for $X = 1$ is $\tau_1 = b + a/ln\ 2$.

In the special case of the Rasch (1961) model, the item parameter a may be absorbed into the latent trait θ and a set to = 1. In this case, the MI score level locations are $\tau_0 = b - .693$ and $\tau_1 = b + .693$.

Implication for Proficiency Level Description

In many situations (such as NAEP), items are selected to represent what students can do or know at predetermined values of the proficiency scale (e.g., the latent trait) so that these points can be described. It seems apparent from the nature of these descriptions that the focus is not the overall location of the item (for the correct and incorrect responses are equally likely). Rather, the focus is the correct score $X = 1$, which signifies that the student can do or knows what is expected on the item. If this presumption is reasonable, then the items to be selected at a given proficiency level are those with a *p*-value of 2/3 or roughly 67%. This observation seems consistent with current Educational Testing Service (ETS)/NAEP practices in selecting items for proficiency level descriptions.

To Locate the Item at b or τ_1

A number of standard-setting procedures focus on what students can do or know of the situation portrayed by various items. It appears then that attention should be paid to the value τ_1, which represents the location of the correct response $X = 1$. Note that the use of τ_1 will result in a higher standard (or passing score) than the traditional reliance on the item location b.

LOCATING THE SCORE LEVELS OF A PARTIAL CREDIT ITEM

The Generalized Partial Credit Model (GPC)

Consider a polytomous (or polychotomous) item in which the score X is equal to 0, 1, . . . , m and, thus, has $m+1$ ordered categories or score levels. At the value of θ for the latent trait, the probability of the score level k is written as

$$P(X = k \mid \theta) = P_k(\theta) = exp\ [a_k(\theta - b_k)]/D(\theta), \qquad (29)$$

where $D(\theta) = \sum_{j=0}^{m} exp[a_j(\theta - b_j)]$. The a_k and b_k are often referred to as item parameters, and the a_ks are (strictly) positive except for $k = 0$, at which both a_0 and b_0 are zero. In addition, when the categories 0, 1, . . ., m are ordered, the a_k must satisfy the inequalities $a_{k-1} < a_k$ so that the ratio $P_k(\theta)/P_{k-1}(\theta)$ will be an increasing function of θ and the density $P_k(\theta)$ has the monotone likelihood ratio property.

This GPC model has been studied extensively by Muraki (1992) and used to analyze polytomous items for large-scale assessment programs such as NAEP and MSPAP (1992). Note that the GPC model belongs to the family of models proposed by Bock (1972) for analysis of items with multiple nominal categories.

If $a_k = k$ for nonzero integer k, then the GPC model coincides with the one formulated by Masters (1982) on the basis of the Rasch model for binary items. The Masters' step difficulties δ_k satisfy the equalities

$$b_k = [\sum_{j=0}^{k} \delta_j]/k,\ k = 0, . . ., m. \qquad (30)$$

By extending Masters' derivations to the 2PL, Muraki (1992) obtained a new formulation and refers to it as the GPC model. Note that the item parameters a_k and b_k used in this paper are the (starred) item parameters a_k^* and b_k^* that Muraki uses in his paper (p. 168).

To facilitate the subsequent derivations and unless otherwise stated, all derivatives signified by the apostrophe such as in $P'(\theta)$ indicate a (partial) derivative with respect to θ. In addition, note that

$$P_k'(\theta) = [a_k - A_1(\theta)]\ P_k(\theta), \tag{31}$$

where

$$A_1(\theta) = \sum_{j=0}^{m} a_j P_j(\theta). \tag{32}$$

The quantity $A_1(\theta)$ is actually the first moment (mean) of the random variable Z, which takes the value a_k with the probability of $P_k(\theta)$. For this random variable, denote the second moment as

$$A_2(\theta) = \sum_{j=0}^{m} a_j^2 P_j(\theta). \tag{33}$$

In addition, note that the variance of Z is equal to

$$\sigma_Z^2(\theta) = A_2(\theta) - A_1^2(\theta). \tag{34}$$

The variance is positive unless the probability $P_k(\theta)$ is degenerated at one unique value. All subsequent derivations assume that the variance of Z is positive.

Item Information and Item Location

The information of the polytomous item is defined as

$$I(\theta) = -E_X\ [\delta^2\ ln\ P(X = k \mid \theta)/\delta\theta^2], \tag{35}$$

where δ^2 denotes the second order partial derivative and E_x the expectation of the random variable X, which represents the score levels of the partial credit item. It follows that the information can be written as the following sum:

$$I(\theta) = -\sum_{k=0}^{m} \frac{\delta^2\ ln\ P_k(\theta)}{\delta\theta^2}\ P_k(\theta). \tag{36}$$

Noting that

$$-ln\ P_k(\theta) = ln\ \{\sum_{j=0}^{m}\ exp[a_j(\theta - b_j)]\} - \sum_{j=0}^{k} a_j(\theta - b_j), \tag{37}$$

it may be verified that

$$\frac{\delta}{\delta\theta}\ [-ln\ P_k(\theta)] = A_1(\theta) - \sum_{j=0}^{k} a_k \tag{38}$$

and

$$\frac{\delta^2}{\delta^2\theta}\,[-ln\ P_k(\theta)] = A_2(\theta) - A_1^2(\theta) = \sigma_Z^2(\theta). \tag{39}$$

Hence, the item information is equal to

$$I(\theta) = \sum_{k=0}^{m} \sigma_Z^2(\theta)\ P_k(\theta) = \sigma_Z^2(\theta). \tag{40}$$

Now as $\theta \rightarrow +\infty$, $P_m\ (\theta) \rightarrow 1$, and as $\theta \rightarrow -\infty$, $P_0\ (\theta) \rightarrow 1$. Hence, as θ goes to infinity, the item information $I(\theta)$ goes to zero. Since $I(\theta)$ is a continuous function of θ, there exists a finite value τ at which $I(\theta)$ is maximized. This quantity τ is called the MI location of the item on the latent trait θ. The location τ, of course, is a function of the item parameters a_k and $_{bk}$, $k = 1, \ldots, m$.

Score Level Information and Location

The score level information is defined as

$$I_k(\theta) = I(\theta)\ P_k(\theta),\ k = 0, 1, \ldots, m. \tag{41}$$

As $\theta \rightarrow \pm\infty$, $I_k(\theta) \rightarrow 0$. With $I_k\ (\theta)$ being continuous, there exists a finite value τ_k at which the score level information $I_k(\theta)$ is maximized. Here, τ_k is the MI location of the score level k.

Theorem 1

Under the GPC model, the MI score level locations τ_k form a strictly increasing sequence. In other words, $\tau_0 < \tau_1 < \ldots < \tau_m$.

Proof

Given

$$I_k(\theta) = I(\theta)\ P_k(\theta) \tag{42}$$

$$\text{and } I_{k-1}(\theta) = I(\theta)\ P_{k-1}(\theta). \tag{43}$$

By noting that

$$P_k(\theta)/P_{k-1}(\theta) = exp\ (u_k\theta - v_k), \tag{44}$$

where

$$u_k = a_k - a_{k-1} \text{ and } v_k = a_k b_k - a_{k-1} b_{k-1}, \tag{45}$$

it follows that

$$I_k(\theta) = I_{k-1}(\theta)\ exp\ (u_k\theta - v_k). \tag{46}$$

Taking the derivative with respect to θ,

$$I_k'(\theta) = [I_{k-1}'(\theta) + u_k I_{k-1}'(\theta)]\ exp\ (u_k\theta - v_k). \tag{47}$$

At the MI score location $\theta = \tau_{k-1}$, the derivative $I_{k-1}(\theta) = 0$ and changes sign from positive to negative.

At this point,

$$I_k'(\tau_{k-1}) = u_k I_{k-1}(\tau_{k-1})\ exp\ (u_k\tau_{k-1} - v_{k-1}) \tag{48}$$

and is positive because $u_k > 0$. Hence, as θ passes through τ_{k-1}, the derivative $I_k'(\theta)$ is still positive and the information $I_k(\theta)$ is still increasing. So, if $I_k(\theta)$ is maximized at the MI score level location τ_k, then τ_k must be larger than the MI score level location τ_{k-1}.

Theorem 2

Under the GPC model, the MI item location τ satisfies the inequalities $\tau_0 < \tau < \tau_m$.

Proof

Given

$$I_k(\theta) = I(\theta)\ P_k(\theta) \tag{49}$$

so that

$$\begin{aligned} I_k'(\theta) &= I'(\theta)\ P_k(\theta) + I(\theta)\ P_k'(\theta) \\ &= I'(\theta)\ P_k(\theta) + I(\theta)\ P_k(\theta)[a_k - A_1(\theta)]. \end{aligned} \tag{50}$$

At the MI item location $\theta = \tau$, $I'(\tau) = 0$, and

$$I_k'(\tau) = I(\tau)\ P_k(\tau)[a_k - A_1(\tau)]. \tag{51}$$

Since $a_0 < A_1(\theta) < a_m$ for all θ, it follows that $I_0'(\tau) < 0$ and $I_m'(\tau) > 0$. When θ increases from the MI score level τ_0, the derivative $I_0(\theta)$ vanishes and turns negative. Hence, $\tau_0 < \tau$. Likewise, when θ decreases from the MI score level τ_m, the derivative $I_m(\theta)$ vanishes and turns positive. Hence, $\tau < \tau_m$.

BOUNDARY BETWEEN SCORE LEVELS

Preliminary Considerations

For a 2PL binary item, the item location b is the ability θ at which an examinee has an equal chance of getting the incorrect response ($X = 0$) and the correct response ($X = 1$). It is the place on the latent trait where the curves for $P(\theta)$ and $Q(\theta)$ intersect. (These curves are also referred to as trace lines.) In some way, b represents the boundary between the scores $X = 0$ and $X = 1$.

Note that for a GPC item, the points on the latent trait where the trace lines intersect may not form an increasing sequence and, therefore, cannot be used as boundaries for the score levels. This section presents a class of boundaries that form an increasing sequence with respect to the score levels in a partial credit item.

Consider again a GPC item with scores $k = 0, 1, \ldots, m$ and with item parameters a_k and b_k. Denote the cumulative probability of having a score of k or less as $G_k(\theta)$. Then

$$G_k(\theta) = \sum_{i=0}^{k} P_i(\theta). \tag{52}$$

To prove that, for a fixed score level k, $G_k(\theta)$ is a strictly decreasing function of θ and that $G_k(\theta)$ decreases from 1 to 0 as θ increases from $-\infty$ to $+\infty$, denote Z_1 and Z_2 as

$$Z_1 = \sum_{i=0}^{k} exp\ [a_i(\theta - b_i)] \tag{53}$$

and

$$Z_2 = \sum_{i=k+1}^{m} exp\ [a_i(\theta - b_i)]. \tag{54}$$

Then it may be verified that

$$G_k(\theta) = Z_1/(Z_1 + Z_2). \tag{55}$$

The derivative of $G_k(\theta)$ with respect to θ can be written as

$$G_k{}'(\theta) = (Z_1{}'Z_2 - Z_1Z_2{}')/(Z_1 + Z_2)^2, \tag{56}$$

where

$$Z_1' = \sum_{i=0}^{k} a_i\ exp\ [a_i(\theta) - b_i)] \tag{57}$$

and

$$Z_2' = \sum_{i=k+1}^{m} a_i\ exp\ [a_i\ (\theta) - b_i)]. \tag{58}$$

Recall that the sequence a_i, $i = 0, 1, \ldots, m$ is strictly increasing. Hence, it may be verified that

$$Z_1' < a_k Z_1 \tag{59}$$

and

$$Z_2' > a_{k+1} Z_2, \tag{60}$$

so that

$$Z_1' Z_2 - Z_1 Z_2' < (a_k - a_{k+1})\, Z_1 Z_2. \tag{61}$$

With $a_k < a_{k+1}$, it follows that the derivative of $G_k(\theta)$ with respect to θ is negative. Thus, the function $G_k(\theta)$ is a strictly decreasing function of θ. In addition, as θ goes to $-\infty$, all the exponential components in this function go to zero; hence, $G_k(\theta)$ will go to 1. Likewise, when θ tends to $+\infty$, the exponential component $exp[a_m(\theta - b_m)]$ in the denominator of $G_k(\theta)$ also goes to $+\infty$ with a dominant order, making the function $G_k(\theta)$ go to zero.

Definition

The boundary between the score levels k and $k+1$ is defined as the value λ_k, at which the cumulative probability $G_k(\lambda_k)$ is equal to 1/2. It is the place on the latent trait where the probability of having the score k or less is equal to the probability of having a score of $k+1$ or more. It is also the place where the combined information of the scores $0, \ldots, k$ is equal to the combined information of the scores $k+1, \ldots, m$.

For a GPC item with maximum score m, there are m boundaries $\lambda_1, \ldots, \lambda_m$. This paper shows that these boundaries form a strictly increasing sequence.

Theorem 3

Under the GPC model, the boundaries λ_1 form a strictly increasing sequence. In other words, $\lambda_1 < \lambda_2 < \ldots < \lambda_m$.

Proof

At the boundary λ_k, $G_k(\lambda_k) = 1/2$. It follows that

$$G_{k+1}(\lambda_k) - 1/2 = P_{k+1}(\lambda_k) > 0. \tag{62}$$

At the next boundary λ_{k+1},

$$G_{k+1}(\lambda_{k+1}) - 1/2 = 0. \tag{63}$$

With the function $G_{k+1}(\theta) - 1/2$ being decreasing, it follows that λ_k must be smaller than λ_{k+1}.

Relationship Between Score Level Location and Boundaries

For a 2PL binary item, $\lambda_1 = b$ and, therefore, is smaller than the location τ_1 of the score $X = 1$. Usually, however, this type of inequality between the score levels τ_i and boundaries λ_i cannot be generalized to a partial credit item. Following is a counter-example.

Counter-Example

Consider the Rasch/Masters partial credit item with $n = 3$; $a_1 = 1$, $a_2 = 2$, $a_3 = 3$; and $b_1 = 0.059$, $b_2 = 0.207$, $b_3 = 0.338$. (The Masters's step difficulties δ are 0.059, 0.355, and 0.599.) The MI locations for the score levels $X = 0, 1, 2$, and 3 are $\tau_0 = -0.35$, $\tau_1 = 0.14$, $\tau_2 = 0.55$, $\tau_3 = 1.03$. The boundaries are $\lambda_1 = -0.42$, $\lambda_2 = 0.35$, and $\lambda_3 = 1.10$. Note that the boundary λ_0 is not situated between the MI locations for the two score levels $X = 0$ and $X = 1$. Note also that the boundary λ_3 is not situated between the MI locations for the two score levels $X = 2$ and $X = 3$.

For most real-life performance items that the author is aware of, however, it is fairly common for the boundary λ_k to be situated between the two MI locations τ_{k-1} and τ_k.

Score Level Location or Boundary?

As in the case of binary items, the use of the MI score level location τ seems to make sense. But the use of the τ score level location usually will result in higher cut scores than the use of the λ boundaries.

ADJUSTMENT OF CUT SCORES

Decision-Theoretic Considerations

In some instances, cut scores may have to be revised to reflect new emphasis on detecting nonadequate students who truly need instructional help. For example, if greater emphasis is put on identifying these students, then it becomes desirable to adjust the passing score upward in order to increase the chance of finding these students.

As in the second section of this paper, denote Q the ratio reflecting the importance of passing an adequate student to the importance of failing a nonadequate student. (For example, if passing an adequate student is less important than failing a nonadequate student, the ratio Q should be less than 1.) In addition, let z be the $100/(1+Q)$ percentile of the normal distribution. Now let c_1 be the cut score set at the priority ratio Q_1 at which the normal percentile is z_1. If the priority ratio is now set at Q_2 at which the normal percentile is z_2, then the new cut score is given as $c_2 = c_1 + (z_2 - z_1) \times \text{SEM}$, where SEM is the standard error of measurement for the group to which the cut scores are appropriate. (Huynh, 1976, 1985)

An Illustration

As an illustration, let $c_1 = 28$ and $Q_1 = 4$. Then $z_1 = -0.84$. Now let the new priority ratio be $Q_2 = 1/2$, at which $z_2 = 0.44$. With an SEM of 1.8, for example, the new cut score now becomes $c_2 = 28 + (0.44 + 0.88) \times 1.8 = 30.3$. If only integer cut scores are used, then the new passing score is to be set at 31.

Assessment of the Priority Ratio Q

In a number of situations it might be possible to set the priority ratio to reflect the purposes of the testing program. For example, if an examinee is given four chances to pass a test in order to be awarded a teaching certificate, then it seems reasonable to set the priority ratio Q at 1/4. Information about the computation of error rates when retaking a test is permitted may be found in Huynh (1990).

Decision Versus Classification

Large-scale assessment programs like NAEP have the purpose of classifying students in various score categories (or achievement levels). For these programs, it appears that the false-positive errors and the false-negative errors are equally important. As a consequence, the priority Q needs to take the value of 1. At this ratio, the z value is zero and the SEM of the test does not enter into the expression for the cut score. Hence, from a decision-theoretic point of view, it seems inconsistent to use the SEM as a basis for making adjustments in cut scores.

SOME PRACTICAL CONSIDERATIONS

Local Item Dependence

Inherent in standard-setting procedures based on judgments on the test items is the assumption that judgments passed on one item do not influence the judgments passed on the other items. This condition implies that the items are (locally) independent from one another. The condition appears easier to meet for multiple-choice items than for constructed-response items based on a common theme such as a long (authentic) reading passage or a mathematical problem. Some details about the contextual characteristics of local item dependency for assessment items of the MSPAP may be found in a study conducted by Ferrara, Huynh, and Baghi (1994) for reading and mathematics items.

Number of Score Levels

It appears that judgments are easier to make for assessment items with fewer score levels (or categories). Working within the Rasch/Masters family for partial credit items, Huynh (1994a, 1994b) proved that all partial credit items can be written as the sum of a locally independent set of binary items (with score levels of 0 and 1) and indecomposible trinary items (with score levels of 0, 1, and 2). It follows that, from a psychometric point of view, it is sufficient to write performance assessment items with no more than three score levels.

The Rose Hill[2] Paradox

In a multiyear assessment program, it is often necessary to field-test new assessment items every year or to use an old test form for linking purposes. Under well-controlled situations (Huynh & Ferrara, 1994), alternate forms of performance assessments can be equated with reasonable results. In a number of situations, however, the degree to which students are familiar with the old form and the level of motivation (or lack of it) that they display on the field-test items can sometimes affect the data on the new test administration in an unexpected manner.

The following example is within the general framework of common item equating and involves the field-testing of new items after the administration of the "operational" test form. Suppose that the examinees know that their scores on the operational test form count, but their scores on the field-test form do not. If lack of motivation on the field-test form leads to subpar performance, then the new items might (statistically) appear harder than they actually are. In this case, it is likely that standards on the old operational test form will be equated to cut scores that are lower than expected on the new operational test form. As a consequence, the use of the new operational form, along with the equated cut scores, would result in an increasing number of passing students, even if student performance does not change from year to year.

---

[2] Rose Hill is a long and steep hill near the University of South Carolina Columbia campus. It is often used by local marathon runners as a route for long (and supposedly easy) training runs.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika, 37,* 29-51.

Ferrara, S., Huynh, H., & Baghi, H. (1994). Contextual characteristics of locally dependent open-ended item clusters in a large-scale performance assessment program. Manuscript paper submitted for publication.

Huynh, H. (1976). Statistical consideration of mastery scores. *Psychometrika, 41,* 65-78.

Huynh, H. (1985). Assessing of mastery of basic skills through summative testing. In D. U. Levine and Associates (Eds.), *Improving student achievement through mastery learning programs* (pp. 185-201). San Francisco: Jossey-Bass.

Huynh, H. (1990). Error rates in competency testing when test retaking is permitted. *Journal of Educational Statistics, 15,* 39-52.

Huynh, H. (1994a). On equivalence between a partial credit item and a set of independent Rasch binary items. *Psychometrika, 59,* 111-119.

Huynh, H. (in press). Decomposition of a Rasch partial credit item into independent binary and indecomposible trinary items. *Psychometrika*.

Huynh, H., & Casteel, J. (1983). *Technical works for basic skills assessment programs* (Final Rep. of Grant NIE-G-90-0129). Columbia, SC: University of South Carolina, College of Education.

Huynh, H., & Ferrara, S. (1994). A comparison of equal percentile and partial credit equatings for performance-based assessments composed of free-responses items. *Journal of Educational Measurement, 31,* 125-141.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). New York: American Council on Education/Macmillan.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Maryland School Performance Assessment Program. (1992). *Final technical report: Maryland School Performance Assessment Program, 1991.* Baltimore: Author.

Maryland School Performance Assessment Program. (1994). *Technical report: Maryland School Performance Assessment Program, 1993.* Baltimore: Author.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology, *In Proceedings of the IV Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 4). Berkeley: University of California Press.

Some Technical Aspects in Standard Setting

Summary of Break-out Session[1]

This session started with a request to have Huynh go over *boundary versus location procedure in locating cut scores*. Huynh presented a 2PL model in a graph that included two slightly overlapping distributions representing two groups with respective means of tau(0) = 0 and tau(1) = 1. He emphasized that what was important about the procedure was the identification of test items at a particular cut point that indicated what a student could do. It was noted that the problem was one of matching the scale difficulty levels with student ability.

Huynh explained that additional problems arose when the complexity of item characteristics increased, e.g., when item changed from binary to polychotomy. He noted that the general practice was to make sure numbers are in the right directions, i.e., in seemingly proper order. Two participants pointed out specific problems in dealing with step difficulty, especially the difficulty of ascertaining meaningfulness when interpretations were made at score levels.

The discussion then focused on *problems associated with employing different methods of assessment that produced different cut scores*. Huynh cited his 1981 study that employed two different methods of assessment but produced essentially the same results. The consistency in the results produced by the two different methods was very reassuring from a technical point of view. He added, however, that context is the most important element in determining the appropriateness of standards.

One participant indicated that in many more situations, different methods produced different results. Another participant expressed the need for standards on standard setting and the desirability of combining different methods. Further discussions indicated that multiple assessments were a good thing and were usually required in order for assessments to be objective. The case of New Hampshire was cited: The results of three sets of assessments were combined and adjustment to the cut scores was made. Eventually, some political decision required adjustments to the standards themselves.

The discussion turned to other specific problems related to standard setting: (a) the distinction between what a student should be able to do and what a student can do, and (b) the need to have sufficient numbers to deal with both dimensions. It was observed that frequently the assessment situation imposes severe constraints on the possible number of items. One specific question related to problems regarding the items pool and the specifications needed to measure what a student can do. Huynh observed that in this context, the length of the description of items was usually poor. Another participant commented on different expectations that often existed among different groups. The group concluded that outside, or external, validations of context and meaning are needed under such circumstances.

---

[1] This is a summary of the break-out discussions for Huynh Huynh's presentation. The session was facilitated by Eugene Johnson (National Center for Education Statistics) and recorded by Shi-Chang Wu (National Center for Education Statistics).

The discussion shifted to *passing scores*. Huynh stated emphatically that passing scores can be set only if it is known what they are to be used for; in other words, only well-defined, conceptual definitions can serve as the guide for identifying appropriate passing scores. The group concurred with this observation. It was acknowledged that that applies to other practices, for example, acceptable performance, referral making, and group performance level, all of which require conceptual definitions (as opposed to technical definitions). This discussion ended with a common recognition of the need to describe what a student can do given a certain score.

The group then focused on the *tester's role in standard setting*, or more generally, *who should set standards*. Various suggestions were presented. Some considered that only professionals with sufficient training in the technical aspects of testing and interpretation should assume the role of standard setters. Others proposed the opposite and suggested that psychometricians should not set standards because (a) standards are not just measurement and technical problems, and (b) other, nontechnical input is necessary. One participant indicated that standard setting is never an empirical matter; a standard is always arbitrary. Huynh then noted that descriptive score levels are deceiving. He referred to his earlier discussion on the importance of translating test scores into what a student could do or could not do. However, he indicated that this was an ideal situation. He cautioned the use of such score levels.

The final segment of this session focused briefly on several other issues. In considering what should be the desirable form for standards, one person indicated that *norm-referenced standards* still dominate testing because the public understands them. The group then considered *different norming methods*. Important attributes noted in the discussion included the characteristics of statistical distribution and the acceptance by users of the test results as reflecting reality. Huynh mentioned the important notion of mathematical approximation (or "continuity") versus precision in discussing the difference of 1PL and 2PL models. He pointed to the difficulty of translating test scores into reality. He called participants' attention to "something out there that we could hang on to set standards"; he meant, among other things, policies, future goals, and consensus among some people. He emphasized the need for national discussions on standards and standard setting.

The session concluded with a broad discussion of two additional topics: the importance of reporting data meaningfully, for example, with close reference to the national goals, and the acknowledgement of the role of the judgmental process in standard setting and, its concomitant, the importance of external reference.

A Conceptual Analysis of Standard Setting in Large-Scale Assessments

Wim. J. van der Linden

Professor of Educational Measurement and Data Analysis,
Faculty of Educational Science and Technology,
University of Twente, Enschede, The Netherlands

ABSTRACT

This paper consists of three parts. The first part analyzes the use of cut scores in large-scale assessments and discusses three different functions of cut scores: (a) They define the qualifications used in assessments; (b) they simplify the reporting of achievement distributions; and (c) they allow for the setting of targets for such distributions. The second part of the paper gives a decision-theoretic alternative to the use of cut scores and shows how each of the three functions identified in the first part can be approached in a way that may reduce some of the feelings of arbitrariness that often accompany standard-setting processes. The third part of the paper formulates criteria for standard-setting methods that can be used to evaluate their results.

It has often been stated that setting standards in large-scale assessments is a process with arbitrary results. The purpose of this paper is to precisely identify the elements of arbitrariness in the standard-setting process, to present an alternative approach to the use of cut scores that may reduce some of the feelings of arbitrariness involved in standard setting, and to provide criteria to distinguish better standards from worse. The main philosophy in this paper is that standard setting will always involve a subjective choice, but that some choices are consistent with empirical data and meet important criteria of rationality whereas others do not.

THREE FUNCTIONS OF STANDARDS

Standard-setting processes in large-scale assessments typically aim at the selection of one or more cut scores on an achievement variable. For simplicity, consider the following analysis of the case of a single cut score. Let θ be a content-referenced-achievement variable on which a cut score has to be selected. Figure 1 contains a graph of the distribution of the examinees in the population on this variable, along with a possible cut score, θ_c, used to distinguish between the two intervals with the qualifications "Satisfactory" and "Unsatisfactory." The selection of this cut score serves three functions:

It defines the qualifications. If the test-score variable is content referenced, in the sense that for each level of the variable the domain of achievements mastered by the examinees is specified, the cut score defines the qualifications in terms of the behavior of examinees. An excellent way to map achievements on a variable is to use response functions from Item Response Theory (IRT). For the domain of achievements represented by the test, these functions specify the types of problems that can be solved with a given probability of success for each possible level of the variable. An early application of response functions to map achievements on a score scale is the *KeyMath* test for the diagnosis of arithmetic skills (Connolly, Nachtman, & Pritchett, 1971), but the origin of the idea to use the content of test items to define a variable underlying a test goes back to Thurstone (1925, Fig. 6).

Thus, contrary to a popular point of view (Kane, 1994), the point taken here is *not* that the meaning of cut scores is defined by what now are generally known as performance standards; it is that the achievements of examinees on the test items provide the scale of test scores with a behavioral interpretation. The role of cut scores is only to link qualifications in assessment studies to these scales and, hence, to their behavioral interpretation. Performance standards are verbal descriptions of achievements that form an important step in the process of specification that leads to the domain of test items represented in the test and selects the cut scores. However, once the domain has been realized and the cut scores selected, performance standards lose their operational meaning. From that point on, conversely, the domain of test items and the cut scores define the empirical meaning of performance standards.

It reports assessment results. If a standard in the form of a cut score is available, it can be used to simplify the reporting of assessment results. The typical statistics used to report such results are estimates of the percentages of examinees, with achievements in the intervals defining the various qualifications. In Figure 1, the shaded area represents the percentage of examinees with the qualification, "Satisfactory."

From a statistical point of view, the process of standard setting is only a form of *data reduction*. Each cut score dicnotomizes the distribution of scores along the full scale. In so doing, information on the relative numbers of examinees on each side of the cut score is retained, but all other information on the shape of the distribution is sacrificed. That the general public finds it difficult to understand the concept of a distribution of scores on a continuous scale is one reason for this data reduction but certainly not a sufficient one. A more important reason has to do with the last function of a cut score identified here.

It sets targets for achievement distributions. The presence of a cut score offers the possibility of setting targets for the outcomes of educational policies in terms of the achievements of the examinees. Typically, these targets are set as upper and lower bounds to the percentages of examinees in the population that meet the various qualifications.

In public discussions of standard setting for large-scale assessments, often no clear distinction is made between the definitions of qualifications and the setting of targets. One reason for this omission might be that high qualifications invariably are perceived as a challenge by some students and teachers and, therefore, tend to serve as de facto standards at an individual level. However, in large-scale assessments, targets are set for *distributions* of achievements, and only bounds to the percentages of students that meet the qualifications can serve as targets. The tendency to automatically perceive qualifications as targets is stronger if the terms used to communicate the qualifications already have an everyday meaning loaded with positive or negative emotions. Looking for terms without an emotional loading is therefore an important activity in standard-setting processes. In this respect, the National Assessment of Educational Progress (NAEP) has done a respectable job in selecting neutral terms such as "Basic," "Proficient," and "Advanced." Nevertheless, it seems a permanent task of assessment specialists to remind the general public that qualifications used in assessment studies have no meaning over and above the behavior of examinees classified by cut scores on achievement scales, and that targets are always to be set for distributions of achievements.

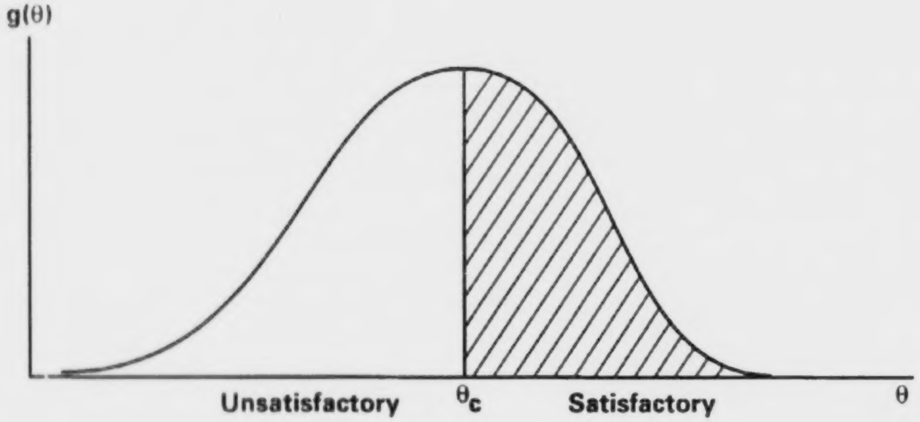


Figure 1. Distribution of students on an achievement variable with cut score θ_c defining two possible qualifications.

It is important to note that a cut score's exact location is not always important when setting targets. For a given distribution of achievements, a lower cut score leads to a higher percentage of examinees above the cut score. As a consequence, different combinations of cut scores and percentages may be met by the same distribution and, hence, imply the same target.

A DECISION-THEORETIC APPROACH TO STANDARD SETTING

The feelings of arbitrariness in standard setting referred to in the introductory section stem from the fact that although cut scores have an "all-or-none" character, their exact location can never be defended sufficiently. Examinees with achievements just below a cut score differ only slightly from those with achievements immediately above this score. However, the personal consequences of this small difference may be tremendous, and it should be no surprise that these examinees can be seen as the victims of arbitrariness in the standard-setting procedure. These feelings of arbitrariness are reinforced if it is noted that the procedure contained random events such as the selection of judges, test items, or experts.

On the other hand, as explained above, the ultimate goal of large-scale assessments is to set targets for achievements. Therefore, neither policymakers nor consumers of education are served well by qualifications phrased in words that have only a vague relation to the achievement variable in question and that cannot be used to set clear-cut targets.

In conclusion, standards are apparently needed that are a little "softer" than the cut scores now generally in use but that nevertheless allow for unequivocal decisions about the quality of education. A decision-theoretic approach to the standard setting problem could be a first step toward this end, although, as will be elucidated below, no procedure will ever remove all subjectivity from standard setting. For each of the three functions of standards described in the previous section, the next section provides a decision-theoretic alternative not based on the use of cut scores on tests.

Defining Qualifications for Large-Scale Assessments

The first function of cut scores could be restated by observing that, at a more formal level, a cut score is nothing but an instance of *a rule to link a set of possible qualifications to an achievement variable*. Other rules are possible and may even be better. A more general rule is exemplified in Figure 2, where the two possible qualifications used in Figure 1 are graphed as a *continuous function* of θ.

Several interpretations of these functions are possible, including the following:

1. A decision-theoretic interpretation would view the two functions as representations of the utilities involved in assigning the qualifications to the various levels of the achievement variable θ. The term "utility" is used here in its technical sense as a measure summarizing all the costs and benefits involved in a decision outcome. Empirical estimates of utility functions are easier to obtain for decisions with immediate personal or institutional consequences, such as selection decisions or mastery decisions for certification purposes, than for large-scale assessments where the consequences of the decision as to which qualification to assign to the achievement distribution of a population of students are indirect and involve only long-term costs and benefits. Empirical estimates of utility functions for selection and mastery-testing problems have been studied by van der Gaag (1990) and by Vrijhof, Mellenbergh, and van den Brink

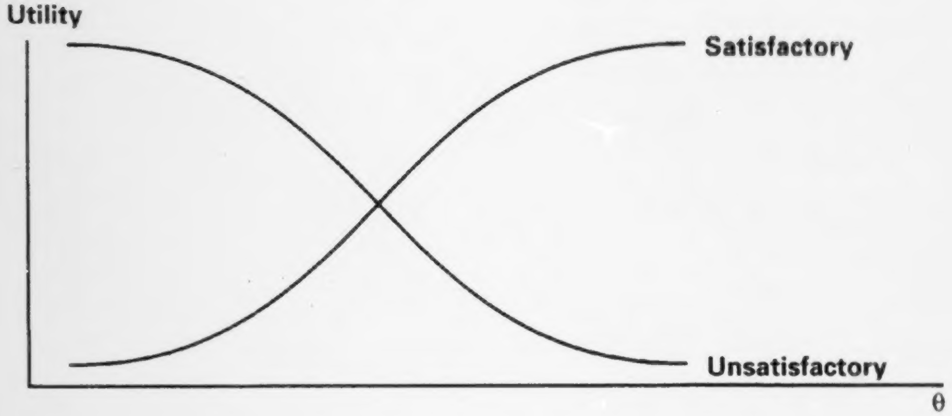


Figure 2. Continuous utility functions used to link the exemplary qualifications to an achievement variable.

(1983). A more general treatment of utility measurement is given in Verheij (1992). Nevertheless, some formal properties of utility functions for large-scale assessments are obvious. For instance, the utility associated with the decision to assign the qualification, "Satisfactory," should be an increasing function of the achievement variable, whereas a decreasing function is needed to model the utility associated with the decision to assign the qualification, "Unsatisfactory." For the conceptual analysis in this paper, these characterizations will do. In actually applying decision theory to large-scale assessment, the shapes of these functions have to be specified further.

2. Large-scale assessment can also be viewed as an attempt to evaluate an achievement distribution on a rating scale. In this view, qualifications such as "Unsatisfactory" and "Satisfactory" correspond with the categories in the rating instrument. In IRT models for the analysis of ratings, the first step is always to scale or locate the categories on the variable measured by the instrument. The results from this stage are then used to rate objects on the same variable. This practice is perfectly consistent with the idea that in large-scale assessment, the first function of standards is to define the possible qualifications in terms of the levels of an achievement variable, and that next these qualifications are used to evaluate the distribution of a population of students on this variable. The response functions in the common IRT models for rating scale analyses with two categories are probability curves with the same decreasing and increasing shapes as the curves in Figure 2. If this probability interpretation is adopted, the curves are each other's complements in the sense that for all levels of the achievement variable, the sums of the two probabilities are equal to one.

3. Suppose a population of judges has used one of the item-centered judgment methods for standard setting--for example, an Angoff (1971) method. However, the task has been to provide probabilities of success not for a "minimally competent" examinee but for an examinee with "satisfactory achievements." The increasing curve in Figure 2 can then represent a smoothed version of the cumulative distribution function of the results from this experiment for the population of judges, while the other curve is the decumulative distribution function for an experiment for examinees with "unsatisfactory achievements."

4. The last interpretation is of an experiment with an examinee-centered judgment method in which one group of examinees is deemed to represent a satisfactory achievement level and the other, an unsatisfactory level. The two curves in Figure 2 could then represent smoothed versions of the cumulative and decumulative distribution functions of the achievements of the two groups, respectively.

Each of these interpretations seems equally possible. However, to remain consistent with the decision-theoretic perspective taken in this paper, the first interpretation is followed.

If more than two qualifications are needed, additional utility curves must be introduced, some of which take a shape different from the ones in Figure 2. Suppose the qualifications "Below Basic," "Basic," "Proficient," and "Advanced" have to be mapped on a NAEP achievement scale. Figure 3 offers a set of utility curves for this problem. The two curves for "Basic" and "Proficient" are bell shaped because assigning these qualifications to very low or very high achievements obviously is a wrong decision, which thus represents a low utility. Because "Proficient" has a more favorable meaning than "Basic,"

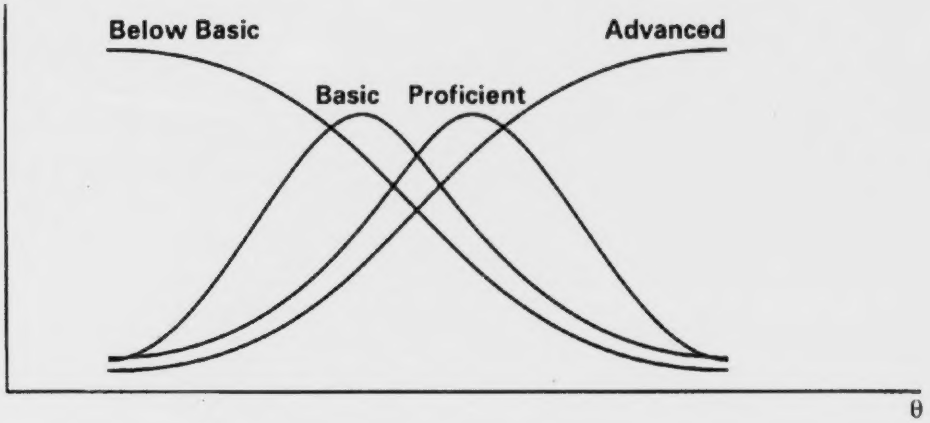


Figure 3. Example of utility functions for the four qualifications used in NAEP.

BEST COPY AVAILABLE

the curve for the former has its location to the right of the latter. The curve for the qualification "Unsatisfactory" decreases, whereas the one for the qualification "Advanced" increases. Note that the usual IRT models for rating scales with more than two categories produce families of response curves with the same shapes.

A Comparison Between Cut Scores and Utility Functions

Cut scores are special cases of the utility functions defined above. Figure 1 follows from Figure 2 if the curves in the latter are taken to represent a *threshold utility function*. Figure 4 illustrates this reduction for a 0-1 threshold utility function. The figure shows that the cut score is the achievement level at which the function for the qualification "Satisfactory" jumps from zero to one while the function for "Unsatisfactory" falls from one to zero. This point of discontinuity, along with the assumed constancy of utility over the intervals to the left and the right of this point, gives cut scores their all-or-nothing character.

It should be noted that the continuous utility functions in Figure 3 have points at which their curves cross, and that these points define intervals for which one of the qualifications has the largest utility. However, these points are not cut scores. Neither do the points at which the functions have maximal values have any decisive meaning. As will be made clear, when assigning qualifications to achievement distributions, it is the *full* shape of the utility functions that counts. The results are, therefore, remarkably robust with respect to the values of these curves at individual points along the scale.

Assigning Qualifications to Population Distributions

Let $g(\theta)$ be the density function representing the distribution of the achievements, Θ, of a population of examinees in an assessment study. The question of how to report assessment results is now reformulated as, "If one qualification has to be assigned to this achievement distribution, which qualification is best?"

An obvious criterion for assigning an optimal qualification is *maximal expected utility*, common in (empirical) Bayesian decision theory. Depending on whether or not $g(\theta)$ is known, various statistical implementations of this criterion are possible. This paper presents the major implementations for the case of two qualifications shown in Figure 2, but the theory applies equally well to cases with any number of qualifications.

$g(\theta)$ known. The following notation is used for the two utility functions: $f_U(\theta)$ (Unsatisfactory) and $f_S(\theta)$ (Satisfactory). The criterion of maximum expected utility indicates that if only one qualification is to be assigned, an optimal choice is one with the largest value for the expected utility. The expected utilities of the two qualifications are calculated with respect to θ as

$$E[f_U(\Theta)] = \int f_U(\theta)g(\theta)d\theta \tag{1}$$

and

$$E[f_S(\Theta)] = \int f_S(\theta)g(\theta)d\theta. \tag{2}$$

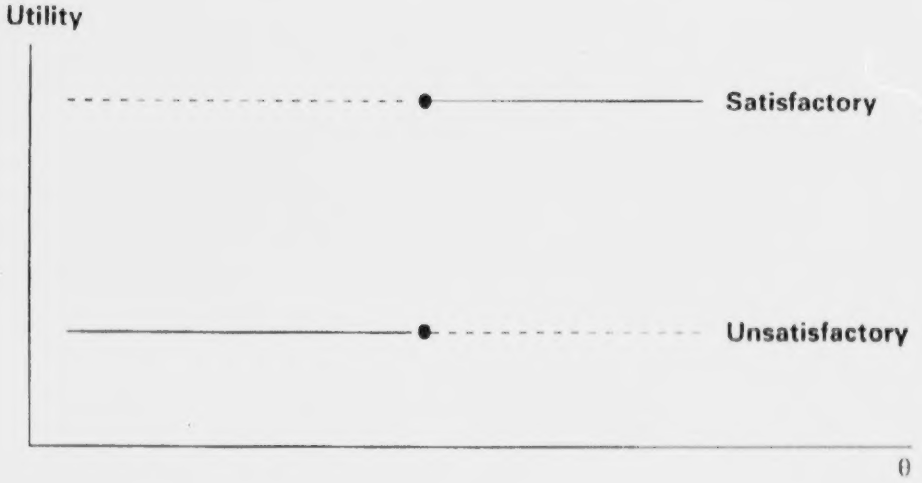

Utility
Satisfactory
Unsatisfactory
θ

Figure 4. A cut score, θ_c, represented as a threshold utility function.

According to the criterion, the qualification "Satisfactory" is an optimal assignment if

$$E[f_S(\Theta)] \quad \geq \quad E[f_U(\Theta)], \tag{3}$$

and the qualification, "Unsatisfactory," is optimal otherwise.

In most assessment studies, interest exists in evaluating the achievements of certain subpopulations. Examples are subpopulations defined by socioeconomic background, by race, or by gender, or subpopulations consisting of, say, the top 5% or bottom 20% of the total population. For each possible subpopulation, the procedure described above can be repeated to determine the best qualification. It is also possible to extract more information from the utility distributions associated with (sub)populations and to report, for example, which qualification is second best; however, this paper does not further pursue these and other obvious refinements.

g(θ) unknown. If gθ is unknown, a Bayesian approach can be used to estimate it based on a distribution of observed test scores. Let $g_o(\theta)$ be a prior for this unknown density and $h(x \mid \theta)$ the density that models the conditional distribution of the observed scores given $\Theta = \theta$. (The case where response data cannot be reduced to a sufficient statistic X will not be addressed here.) From Bayes' theorem, it follows that the posterior density, $k(\theta \mid x)$, is given by

$$k(\theta|x) = \frac{h(x|\theta)\,g_o(\theta)}{\int h(x|\theta)\,g_o(\theta)\,d\theta}. \tag{4}$$

A large range of options is available to choose the prior $g_o(\theta)$ in Equation 4. For example, it is possible to choose a noninformative prior, to use an informative prior based on previous knowledge about $g(\theta)$, to estimate $g(\theta)$ along with the parameters in $f(x \mid \theta)$ in an "empirical" Bayes fashion, or to follow a hierarchical Bayes approach. In addition, collateral empirical information can be used to improve the estimate of the posterior density. More information on each of these approaches is available in standard textbooks on Bayesian statistics.

If an estimate of the posterior density is obtained, it can be used to obtain an estimate of the marginal density $g(\theta)$ through

$$g_1(\theta) = \int k(\theta|x)\,p(x)\,dx, \tag{5}$$

where $p(x)$ is the density of the observed scores. If $g_1(\theta)$ is substituted in Equations 1-3, estimates of the expected utilities for both qualifications are obtained, and the maximum expected utility criterion is now applied to these estimates.

Formulating a Target for an Achievement Distribution

The maximum expected utility criterion indicates what the best qualification for an achievement distribution is but not whether the targets are met. However, if the only decision problem is which qualification to assign, setting a target for the outcome is as simple as choosing the qualification that should be best for the population being assessed. If, in addition, the achievements of subpopulations are assessed, targets have to be set for their distributions, too. For some subpopulations the targets will be the same, but for others it makes sense to set different targets. Some fictitious examples of targets with the format suggested here are:

1. The qualification for the achievements in English of the national populations of students should be "Proficient," with "Advanced" rather than "Basic" as second best qualification.

2. The top 20% of the national population should have achievements in mathematics that qualify as "Advanced."

3. The bottom 10% of the subpopulation of children of first-generation immigrants should have achievements in English with the qualification "Basic."

4. For all possible subpopulations and subjects, the qualifications should not show any differences with respect to gender.

Again, obvious refinements of these suggestions will not be pursued here; the main purpose of this paper is only to outline an alternative to the current practice of large-scale assessment.

STANDARDS FOR STANDARD SETTING

The final part of this paper formulates criteria that can be used to discriminate between better and worse standards. "Standard" is used here as a generic term; there is no need to choose between experiments to establish utility functions and the experiments currently used to select cut scores. Some of the criteria are derived from the statistical literature, whereas others address practical issues or are more empirically oriented. This presentation is only a first attempt. With the growing international interest in the potential of large-scale assessment for enhancing the quality of education, it would be worthwhile to take coordinated action to obtain a more elaborated list.

Explicitness. The criterion of explicitness is not new and, in fact, applies to any research activity or scientific procedure. The criterion stipulates that all steps in a standard-setting experiment should be based on explicit definitions and procedures. First, the motivation of this criterion is communicative. If this criterion is not met, standard setters can never communicate their results in a meaningful way. But technical reasons also exist for this criterion. Without an explicit definition of the steps in the standard-setting procedure, it would never be possible to apply any of the criteria below. For example, without this information, it would be impossible to determine if the procedure was subject to possible inconsistencies, or to replicate the procedure to estimate its statistical properties.

Efficiency. This statistical criterion is defined with respect to the variability of the results from a standard-setting procedure across replications. The lower the variability, the more efficient the procedure.

An important step in designing an experiment to estimate the efficiency of a standard-setting procedure is to determine which aspects of the procedure are allowed to vary across replications and which are not. Generally, those aspects that are allowed to vary are irrelevant and should not introduce any variability in the resulting standards. However, the basic aspects of the method should be kept fixed.

If they nevertheless do vary, then variability in the outcomes is to be expected and is not necessarily a bad thing.

Examples of irrelevant aspects of standard-setting procedures are situational factors such as occasion and location. It would be difficult to accept the possible impact of these factors on the results of the procedure. In other words, the procedure should be efficient with respect to variation on these factors.

However, if a method is based on judgment of the contents of test items, then the question whether items are allowed to vary across replications cannot be answered without considering the scale of the achievement variable. If the achievements are scored on the number-right scale, the properties of test items are not irrelevant, and variability of results due to item sampling is to be expected. Items do differ in their psychometric properties, and a standard-setting procedure that does not reflect the impact of these properties on the achievements of the examinees will even be a bad method. On the other hand, if sampling is from a pool of items calibrated on the same achievement scale, using an IRT model that allows for all of the differences between the items, and if the standard is set on this scale, then the procedure should have high efficiency with respect to item sampling.

It is a common observation that results from standard-setting experiments show variation across methods. This variation is to be expected. Each method instructs its subjects to a different task. At a more general level, as emphasized by the stimuli-organism-response (S-O-R) paradigm in psychology, it holds that for any type of stimuli, responses from subjects depend on the properties of the stimuli as much as on the properties of the subjects. For the standard-setting process, the dependency is as depicted in Figure 5.

The same relation is well known to test theorists who have struggled for decades to find a way to separate the properties of test items from the abilities of the examinees. These theorists now use IRT models to calibrate the properties of the items before they are used as a measurement instrument and use these calibrations to equate test scores from different instruments. A similar development should take place in standard setting. The right approach is not to view variation among methods as error or random noise that is only there to be averaged out, but to calibrate these methods and use the calibrations to equate results from one method to those from another.

By symmetry, the argument above also applies to variability in standard-setting results across judges. This variability is to be expected because judges have their own views about what to expect from education. If judges have worked carefully and meet all the other criteria on the list, then a standard-setting method should allow for differences in views among judges instead of suppressing them as random noise. Obviously, to get practical results, standards set by different judges always have to be combined into a single standard. The point, however, is that the question of what a good standard-setting method is, should be separated from the question of how optimally to combine different standards into a single standard. Decision-theoretic approaches to the latter question also exist. The question is then approached as an instance of the problem of how to best represent a distribution of numbers by a single number, or of how to combine collective choices into a single order of preferences.

Unbiasedness. Generally, a method is unbiased if it produces estimates that, on average, are equal to the true parameter value. The criterion is discussed here because some policymakers and educators seem to believe that true standards exist independently of methods and judges, and that a standard-setting method only should mirror these standards as accurately as possible. This view is not correct.

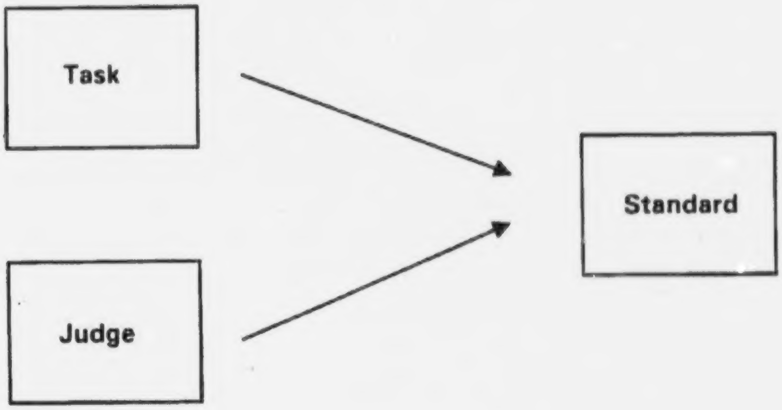


Figure 5. Interaction among judge, task, and standard.

First, standards do not exist without achievement variables, and those variables are not natural quantities but human inventions used to score examinees in tests. Further, as has already been pointed out, judges have different views about what to expect from education, and different views entail different standards. In addition, as was argued in the previous section, even for a single judge, standards do not exist independently of the method used to set them. The correct view is to see standard-setting methods as methods to *set* true standards--not reflect them.

The belief that true standards exist independently brings to mind another classical struggle in test theory--the one with the concept of true score. During the first half of this century, many theorists behaved as if for each examinee there existed a numerical score that exclusively represented the examinee's true ability on the variable measured by the test. While this view implied that some tests were less biased than others, it did not offer any suggestions as to how to estimate bias in tests. The view has been completely abandoned, and currently no test theorist believes in this so-called concept of a *Platonic* true score (Lord & Novick, 1968, sect. 2.9). True scores are now generally defined as expected or average test scores observed across replicated administrations of the same test with the same examinee. In this approach, test scores are unbiased by definition. Standard-setting theorists should learn from this experience and follow the same direction.

Consistency. Although it is customary to speak about data in standard-setting experiments, it is important not to forget that these data are judgmental. Examples of judgments in standard-setting experiments are (a) estimates of subjective probabilities in an Angoff experiment; and (b) ratings of expertise in an experiment with an extreme-group method. However, judgments can be inconsistent in the sense that two or more judgments contradict each other or that their combination contradicts reality. The following three examples illustrate possible inconsistencies:

1. From paired-comparison experiments in scaling, it is known that judges may display intransitivities. Intransitivities would occur in a standard-setting experiment if a judge rated Subject A more proficient than Subject B, B more proficient than C, but A less proficient than C.

2. In an Angoff experiment, judges can specify a high probability of success for a difficult item and a low probability of success for an easy item for the same "borderline examinee." Both probabilities imply standards that can never exist at the same time. Such inconsistencies were frequently observed in an analysis reported in van der Linden (1982).

3. The same type of inconsistency may occur at test level if two tests, A and B, have a high correlation between the scores X_A and X_B because they measure the same variable. If both tests are used to select cut scores in a standard-setting experiment, and if some of the cut scores for Test B cannot be predicted from those for Test A by the regression function of X_B on X_A, then these cut scores are inconsistent.

Note that the last criterion need not hold for test scores measuring *different* achievement variables. For example, the fact that different achievement variables correlate highly for a given curriculum does not imply that if a judge sets a high standard on one variable, the judge should also set a high standard on the other.

A more general interpretation of the notion of consistency for use in evaluation research along the lines of the definition above is given in van der Linden and Zwarts (1995).

Feasibility. This criterion is of a more practical nature and deals with the aspect of standards earlier identified as target setting for achievement distributions. Only a loose description will be given. Nevertheless, the criterion seems to focus on an issue that is on the minds of many critics of standard-setting results. Two definitions of this criterion are possible--one for standards in the form of cut scores and another for standards defined as utility functions:

1. In a cut-score-based approach to large-scale assessment, a target is feasible if the resources are available to meet it.
2. In a decision-theoretic approach to large-scale assessment, a target is feasible if it is based on utility functions that incorporate a realistic estimate of the costs needed to realize the levels of the achievement variable.

It is here that the practical meaning of a decision-theoretic approach becomes fully clear. When following a cut-score-based approach, it is easy to forget the idea of feasibility and to ignore the pains and costs it may take to realize a target. However, an attempt to establish utility functions would directly address issues such as costs of resources and their impact. This fact explains why estimating utility functions is generally more difficult than selecting cut scores.

The notion of feasibility is particularly important if standards have to be set with respect to more than one achievement variable. If the resources are fixed, trade-offs among achievement distributions on different variables is typical. For example, it would not be too difficult for schools to produce high achievements in geography if all other subjects could be dropped. Neglecting such trade-offs may be why standard setters, no matter what their expertise in the pertinent domain of content, often show a tendency to set standards unrealistically high when confronted with the task to address a single achievement variable.

Robustness. Some standard-setting experiments provide their subjects with information on properties of the items or on the ability distributions of reference groups. This information is mainly in the form of statistical estimates. In addition, estimates of item properties may be needed to calculate standards from the judgments by the subjects in the experiment. The criterion of robustness is suggested to deal with the possibility of errors in these estimates.

Generally, a standard is robust if minor changes in the data used in the experiment do not lead to changes in it. Robustness is a welcome property because it indicates that uncertainty about certain relevant aspects of reality is not critical to the results of the experiment. Robustness of standards can be assessed through a series of analyses in which changes are made in the values of the estimates and their effects on the behavior of the standards are ascertained.

It is important to note the similarities and dissimilarities between the criteria of efficiency and robustness. Both criteria are based on the idea of replication. In estimating the efficiency of a standard-setting experiment, apart from possible irrelevant aspects, its procedure is left intact. In a robustness study, the procedure is also replicated exactly, but changes are made in the empirical data presented to the subjects. If these data play a role only in the calculation of the standard from the judgments in the experiment, robustness analysis can take the form of computer simulation in which standards are recalculated from data with a simulated error term.

An example of robustness analysis in an evaluation project based on the outcomes of large-scale assessment is given in van der Linden and Zwarts (1995). In the project, the definition of the standards supposed the presence of an intact item pool that was reduced slightly, however, because of pretesting of the pool. A simulation of several item analyses procedures showed that the effects of item removal on the standards seemed negligible.

Evaluation of Standards

Each of the criteria listed should constrain the choice of standards. A possible sequence of constraints is depicted in Figure 6. The Venn diagram shows the set of all possible outcomes of a standard-setting experiment, not all of which necessarily meet each criterion. An empirical check is needed to demonstrate that the proposed standards belong to the subsets of consistent, efficient, and robust outcomes. In addition, the body of knowledge and insights produced by educational research should indicate that the proposed standards are in the subset of feasible outcomes. An outcome that meets the whole list of criteria is in the shaded area in the diagram.

DISCUSSION

It is surprising that some discussions on standard setting do not pay much attention to lessons learned in the history of test theory. Psychometrically, a standard-setting method is nothing but an instrument to elicit responses from subjects from which an estimate of a quantity is inferred. The same formal description holds for an achievement test, an attitude instrument, or a rating scale. It was suggested earlier that the relation between standard setting and rating is particularly close because both activities have an aspect of evaluation. The idea that standard-setting methods should reflect true or Platonic standards, existing independently of standard-setting methods and subjects, seems to underlie some of the arguments in standard-setting discussions but has not proved to be fruitful in the history of test theory. This idea leads to the belief that properties of standard-setting methods only add random error to the true standard, and that the best method of getting an "unbiased estimate" of this standard is to average out differences among methods. The same idea supports the practice of averaging standards over judges to get rid of their "idiosyncrasies." The question of how to separate the impact of methods from the evaluation of judges is the same as that of how to separate the properties of test items from the abilities of examinees. At a more abstract level, the problem is known as the problem of parameter separability or the problem of nuisance parameters in statistics. Standard setters should look deeper into the analogies between problems in standard setting and problems in test theory or scaling and profit from the solutions reached there.

This paper began with the observation that in setting standards for large-scale assessment, interaction of empirical information and subjective choice is typical. Figure 5 shows how these inputs interact in an ideal standard-setting process. Each of the subsets in the diagram is defined by *empirical information* from research for which rigorous methodology exists. However, it is left to the *subjective evaluation* of the judges to prefer one possible outcome in the intersection of the subsets over the others.

Finally, this paper claims that the feelings of arbitrariness typical of standard-setting processes could be reduced further if the definition of standards included the class of continuous utility functions. As demonstrated in Figure 4, this step would mean only that the rigid form of the threshold utility function

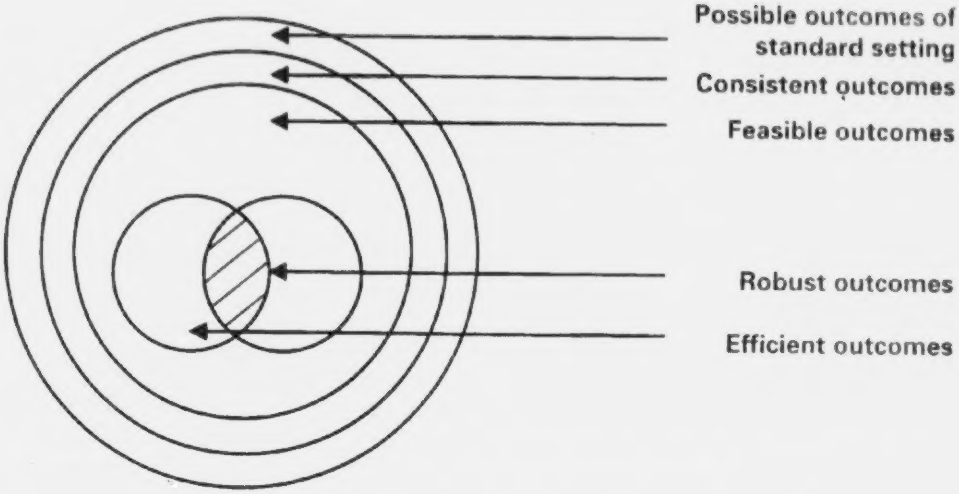

Possible outcomes of standard setting
Consistent outcomes
Feasible outcomes
Robust outcomes
Efficient outcomes

Figure 6. Venn diagram with possible outcomes of a standard-setting experiment.

was relaxed to allow utility functions with such continuous shapes as those in Figures 2 and 3. The impression exists that threshold functions such as those shown in Figure 4 are not obvious candidates for use in assessment studies, and that if a threshold function is met, the precise location of the jump can seldom be defended convincingly. Because assessment results are extremely sensitive to the location of this jump, a cut-score-based approach is bound to be dogged by criticism and feelings of arbitrariness. As has been pointed out, a decision-theoretic approach with continuous utility functions has no individual points on the achievement scale with a dramatic impact on the results of the assessment. However, although establishing the realistic shapes of such utility functions is not a sinecure, the criterion of feasibility suggests that an attempt should be made to establish them.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Connolly, A. J., Nachtman, W., & Pritchett, E. M. (1971). *KeyMath diagnostic arithmetic test*. Circle Pines, MN: American Guidance Service.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64,* 425-461.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16, 433-451.

van der Gaag, N. L. (1990). *Empirische utiliteiten voor psychometrische beslissingen* [Empirical utilities for psychometric decisions]. Unpublished doctoral dissertation, University of Amsterdam, Department of Psychology, The Netherlands.

van der Linden, W. J. (1982). A latent trait method for determining intrajudge inconsistencies in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement, 19,* 295-308.

van der Linden, W. J., & Zwarts, M. A. (1995). Robustness of judgments in evaluation research. *Tijdschrift voor Onderwijsresearch, 20,* 13-27.

Verheij, H. (1992). *Measuring utility: A psychometric approach*. Unpublished doctoral dissertation, University of Amsterdam, Department of Psychology, The Netherlands.

Vrijhof, B. J., Mellenbergh, G. J., & van den Brink, W. P. (1983). Assessing and studying utility functions in psychometric theory. *Applied Psychological Measurement, 7,* 341-357.

A Conceptual Analysis of Standard Setting in Large-Scale Assessments

Summary of Break-out Session[1]

Participants began with general discussions on: (a) the different purposes of standard-setting activities, (b) the extent to which standard setting is arbitrary and/or subjective, (c) criteria for evaluating standard setting and how these change, and (d) the precision and impact of cut scores. It was recommended that good standards should be prefaced with a statement of purpose and should include information relative to the population of test takers to which they apply.

The group reiterated that there is a need to set standards. This discussion included whether standards should be horizontal or vertical, descriptive, judgmental, or hierarchical. Additionally, the group explored whether there should be common standards for all students.

In response to a question of how sensitive a test should be, van der Linden underscored that *efficiency* is a function of variability across standards. He elaborated by saying that low variability indicates robustness (the degree to which the test can be replicated) and determines relevance or irrelevance. The participants indicated that there is a need for more experiments to support efficiency.

One participant asked how to apply *replication* to the National Assessment of Educational Progress (NAEP). Van der Linden explained that replication means repeating the test using the same sample to estimate efficiency and that some parameters hold better than others. He further explained that varying the judges does not constitute replication because each judge has his or her own standards. The judge is a relevant factor.

In discussing arbitrary versus subjective, van der Linden explained the difference in terms of conceptual standards and the true score. He further indicated that in terms of the implications for practice, standard setting is not an important issue for most school-given exams. There is no standard setting here, only equating. To set standards one must scale categories or items and use these categories as measures. There is an explicit empirical meaning to scaling. To assess means to *scale a distribution*. Scaling, measurement, and establishing targets are all needed for assessment.

One participant questioned the appropriateness and/or correctness of the fact that standards reflect different values. The question was posed whether a different approach is needed and if there are legal implications. Van der Linden said that the same approach can be used with either a utility function or a scale distribution. Another participant commented that with no utility function, standards can be set as high as possible. This participant indicated that it is difficult to formulate a utility function for groups.

Van der Linden agreed that utility functions are not appropriate for a large-scale assessment. He suggested that it is necessary, first, to define the terminology of achievement variables, then to set the target. The cut score does not relate directly to curves such as basic and proficient; the difference between basic and proficient is not always sharply defined.

---

[1] This is a summary of the break-out discussions for Wim J. van der Linden's presentation. The session was facilitated by Mark Reckase (American College Testing) and recorded by Kristin Keogh (National Center for Education Statistics).

Some participants commented that, nationally, standards are not always enforced. Another participant contested the generalization that employers are more likely to hire persons who have achieved a high educational standard than persons who have not achieved this high standard. At this juncture, the group discussed whether standards are written for the students or the employers. Van der Linden clarified that standards can be used as a qualification or as a target. In the case of employment, standards are used as a target.

The continuing discussion included a query on the *arbitrariness of standard setting* as related to the political environment and the consequences of standard setting. It was suggested that the political consequences can themselves become the target. While there was a reiteration of the fact that standard setting is an objective process, one participant expressed the view that NAEP itself is associated with a political agenda--e.g., to increase the numbers of students who meet the standards.

The discussion then focused on the *consequences of NAEP standard setting*, particularly the ability to compare American students with those of other nations. One participant suggested that the NAEP scale should have a utility function for the workplace, related to real-world uses such as skill in the workplace. The group was reminded that NAEP has benchmarks and that understanding the meaning of the reported benchmarks is a real issue. For example, many people have difficulty understanding what "2% of students are 'Advanced'" means. It was suggested that the threshold for "Advanced" be revised, because there is a lot of measurement error around 2%. It was also observed that there are unintended consequences at the state level; some states set standards differently from NAEP because of money implications.

There were various suggestions relating to *who should be setting the standards*. It was recommended that a representative model, instead of an expert model, be used, and that standards without accompanying items that serve as exemplars of the various proficiency levels are not useful. There was no agreement among the participants on the use of teachers as panelists. Some criticized the use of teachers, arguing that teachers are biased; others supported the use of teachers, arguing, first, that many teachers are qualified, and, second, that teachers are practiced because they set standards in the classroom. It was further suggested that students can set the standards. Van der Linden interjected that anybody can set standards, provided that they are explicit about the standards. He suggested, however, that the group consider who has the authority to set standards.

Examinee-Centered vs. Task-Centered Standard Setting

Michael Kane

Professor, Department of Kinesiology, University of Wisconsin at Madison

ABSTRACT

Alternate approaches to standard setting cannot be compared in terms of accuracy because there is no way of unambiguously determining where the standard and the corresponding threshold score should be. This paper examines three general criteria for evaluating examinee-centered and task-centered standard-setting methods in different contexts: consistency with the model of achievement underlying test interpretation, the level of demand placed on judges, and technical adequacy. As one might expect, neither method comes out a clear winner in all cases; each has strengths and weaknesses that make it relatively more promising for some standard-setting tasks and in some contexts.

Standard setting changes the interpretation of test scores in several important ways. First, it adds an explicit decision rule involving one or more threshold scores (or "cut scores," or "passing scores") to the basic interpretation of the score scale. The decision rule assigns each examinee to one of several categories depending on how the examinee's score compares to the threshold scores. So a layer of interpretation based on category membership is added to the underlying interpretation based on the scores.

Second, the use of threshold scores to define categories tends to de-emphasize certain information. Typically, the interpretation of the basic-score scale gives roughly equal emphasis to differences in scores, wherever they happen to fall on the scale. To the extent that more emphasis is given to category membership, less emphasis is given to differences within categories.

Assuming that the threshold scores are not selected in a completely arbitrary manner, they must be chosen to represent some intended distinction between different levels of performance. The cluster of intended distinctions between passing examinees (those who have achieved the desired level of performance) and failing examinees (those who have not achieved the desired level of performance) defines a *performance standard*. The *threshold score* is the point on the score scale that is used operationally to categorize examinees. As these terms are used here, the threshold score is a point on the score scale, and the performance standard is the level of achievement associated with that point. The performance standard provides the interpretation assigned to the threshold score.

The interpretation of assessment results in terms of category membership involves at least two assumptions, a *descriptive assumption* and a *policy assumption* (Kane, 1994). The descriptive assumption claims that the threshold scores correspond to specified performance standards in the sense that examinees with scores above a threshold score are likely to meet the corresponding standard of performance, and examinees with scores below the threshold score are not likely to meet the standard. The policy assumption claims that the performance standards associated with threshold scores are appropriate for the intended purpose.

The introduction of performance standards adds the score-based decisions and the consequences of these decisions to the basic interpretation of test scores. The new elements in this extended, standards-based interpretation require additional evidence of validity. In particular, they require support for the additional inferences, decisions, and possible consequences that have been added to the basic interpretation of the test scores (Messick, 1989). In order to develop a convincing case for the validity of the extended interpretation, it is necessary to choose a standard-setting method judiciously and to generate evidential support for the assumptions embedded in the standards-based interpretation.

STANDARD-SETTING METHODS

There are many ways of categorizing standard-setting methods, but the distinction drawn by Jaeger (1989) between test-centered methods and examinee-centered methods seems particularly fundamental. (Note: Jaeger talks about "test-centered models" and "examinee-centered models," but it will be more convenient here to talk about "test-centered methods" and "examinee-centered methods.") The test-centered methods are based on judgments about the test. The examinee-centered methods are based on judgments about the performances of individual examinees. All standard setting is based on judgments.

In the test-centered methods, judges review the items or tasks in the test and decide on the level of performance on each of these items or tasks that will be considered just adequate to meet the performance standard. The judgments about the items in the test are then aggregated in some way to arrive at overall threshold scores. For example, in the Angoff (1971) procedure, judges are asked to envision a borderline examinee and to choose a *minimum pass level*, or MPL, representing the level of performance to be expected of borderline examinees on each item. The MPLs are averaged over judges to get the item MPL, and all the item MPLs in the test are added together to get a passing score. The procedures proposed by Angoff (1971), as well as those proposed by Nedelsky (1954), Ebel (1972), and Jaeger (1982), are test-centered methods in the sense that they require the judges to rate test items or tasks, rather than examinee performances.

In the examinee-centered methods, judges categorize examinees based on some sample of examinee performance, and the passing score is set by identifying a point on the score scale that would be most consistent with these decisions. In the borderline-group method (Livingston & Zieky, 1982), judges identify individuals who are borderline, in the sense that their level of achievement is at or near the performance standard. The median score, or some other index of central tendency, for this group of borderline examinees can then be used as the passing score. In the contrasting-groups method (Livingston & Zieky, 1982), the judges categorize a sample of examinees into two groups, those judged to have met the requirements in the performance standard and those judged not to have met this standard. The passing score is chosen so that it discriminates as well as possible between the upper contrasting group and the lower contrasting group.

The National Assessment Governing Board's (NAGB) effort to establish standards defining three levels of achievement--"Basic," "Proficient," and "Advanced"--has relied mainly on test-centered methods, in particular a version of the Angoff method. The National Academy of Education (NAE) report (Shepard, Glaser, Linn, & Bohrnstedt, 1993) concluded that "the Angoff procedure is fundamentally flawed because it depends on cognitive judgments that are virtually impossible to make" (p. 77), and reiterated this claim in the Executive Summary (p. xxii). The NAE panel also suggested that the contrasting-groups method" appears to be superior to the modified Angoff procedure," because the examinee-centered methods "are based on judgments about *actual* individual students rather than hypothetical populations borderline examinees" (p. 89). However, the objections raised by the NAE report against the Angoff method are not universally accepted (Cizek, 1993; Kane, 1993; Mullins &

Green, 1994) and, therefore, the issue of whether an examinee-centered or a test-centered method would be more appropriate for National Assessment of Educational Progress (NAEP) deserves attention.

These two general approaches to standard setting have both been around for some time, but until recently, the examinee-centered methods have not been given much attention. Most high-stakes testing programs used one or another of the test-centered methods, often some variant of the Angoff procedure, and most published research on standard setting has focused on the test-centered methods. However, this situation is changing, with more attention being given to the examinee-centered methods.

The task of comparing the examinee-centered and test-centered standard-setting methods and drawing firm conclusions about which approach should be preferred is quite difficult, because there is no external criterion against which the two methods can be compared. There is no third method that can be used as the gold standard.

Nevertheless, the task of comparing the examinee-centered and task-centered standard-setting methods and drawing some conclusions about which approach should be preferred is not hopeless. Even though the two approaches cannot be compared directly to a criterion, it is possible to compare the methods in terms of how well they meet general criteria of: (a) congruence with the cognitive model underlying the testing program; (b) practicality, or ease of implementation, particularly in terms of the demands placed on judges; and (c) technical characteristics of the results. These three criteria are used here to evaluate the test-centered and the examinee-centered methods in terms of how well they satisfy the criteria in different contexts. The discussion will focus on two kinds of tests: objective tests, consisting of a large number of specific questions requiring relatively short answers, and performance tests, involving a few, more complex tasks, each of which calls for a relatively extended response.

MODELS OF LEARNING AND ACHIEVEMENT

The most general level on which the two approaches to standard setting can be compared involves questions about how we view learning and achievement and, therefore, about how we tend to characterize the decisions to be based on threshold scores. It seems reasonable to expect that all aspects of an assessment program, including test development, scoring, standard setting, and reporting of results, should be consistent with the intended interpretation of the results.

The discussion here concentrates on two types of models that figure prominently in current discussions of testing. The two types of models--holistic models and analytic models--actually anchor the ends of a continuum of possible models, in the sense that the operant model in a particular context may incorporate some holistic features and some analytic features. I will argue later that NAEP has both analytic and holistic components.

Holistic Models

Holistic models assume that achievement or skill tends to be highly integrated and, therefore, that the only meaningful way to assess achievement adequately is to observe extended performances, preferably actual samples of the performance of interest. It is considered essential for the scorer to get a sense of the richness and complexity of a student's performance in a realistic context. In terms of its implications for assessment, the essential element in holistic models is the assumption that significant, authentic performances cannot be broken down into a series of small, independent tasks without destroying the essential meaning of the performance.

Holistic models have traditionally been dominant in the arts and in areas like physical education and sports, where there is a strong focus on specific kinds of performance. In these areas, instruction may give some attention to the development and assessment of specific skills (even here there can be an analytic element), but the clear goal is for students to be able to integrate a range of skills in an effective performance. Recently, assessments based on holistic models have become increasingly popular in many areas of assessment, in part because of perceived advantages in validity for performance assessments as measures of desired outcomes (Wiggins, 1993), and in part because of a growing awareness of the impact of assessment on teaching and learning (Crooks, 1988).

Some form of performance testing would seem to be essential if one adopts a holistic view of achievement. That is, if it is the performance as a whole that is of interest, and it does not make much sense to evaluate parts of the performance in isolation, then it would seem appropriate to assess achievement by evaluating complete performances in realistic contexts.

Since the holistic model does not view the use of a set of component tasks as a sensible way to assess overall performance in the activity--even if this set of discrete component tasks is relatively comprehensive--it would not be very sensible under this model to set the standard by rating component tasks or by rating performances on these component tasks. Rather, to be consistent with the holistic model, standard-setting studies should focus on complete performances.

The examinee-centered methods seem to fit a holistic model especially well. These methods typically ask the judges to evaluate each examinee's performance as a whole, rather than using any piecemeal approach. The use of judgments about the examinee's overall performance is clearly consistent with the view that it is the performance as a whole that is meaningful.

Analytic Models

Analytic models assume that achievement can be assessed adequately using relatively small parts or samples of performance as indicators of achievement. The analytic models do not necessarily imply that achievement lacks structure, but they do imply, at least implicitly, that it is possible to draw inferences about the achievement or skill of interest from observations on specific tasks associated with the achievement, even though each of these tasks may constitute a relatively small part of the overall performance.

The analytic model provides a framework in which it makes sense to score an examinee's performance on a large number of short, discrete tasks, and to aggregate these task scores into a single test score representing the examinee's overall level of achievement. Furthermore, given the convenience and efficiency of objective testing, this approach to test design is a very attractive alternative within the context of an analytic model of achievement, unless of course the use of these methods might have negative consequences for teaching and learning (Crooks, 1988).

The test-centered standard-setting methods, in which judges rate individual tasks or items, and the threshold score is derived by adding these item ratings together, make a lot of sense in the context of an analytic model. Assuming that examinees are evaluated in terms of their responses to a number of independent tasks or items, it seems reasonable to focus the standard-setting effort on these separate tasks. Furthermore, if we are going to use a specific scoring rule to combine a student's scores on a number of independent, component tasks into the student's overall score, it seems reasonable to determine the threshold score by using the same scoring rule to combine judgments about the demands inherent in these separate component tasks. The examinee-centered methods, which usually focus on extended performances, are not necessarily inconsistent with an analytic model of achievement, but neither do they seem especially appropriate.

In general, it seems that performance tests and examinee-centered standard setting provide a good fit to holistic models of achievement, and that short answer or objective tests and test-centered standard setting provide a good fit to analytic models. So for example, under the analytic view, it is reasonable to at least consider assessing writing ability using a set of questions dealing with syntax, punctuation, word use, and sentence structure, and/or relatively short samples of writing or editing. By contrast, the holistic view tends to assume that the only way to meaningfully assess writing is to evaluate the person's performance in writing a full essay or story. Observing performance on a large number of discrete, component tasks and trying to draw conclusions about overall achievement does not make much sense under a holistic model.

It is not my intent to argue for or against either of these models as being generally preferable to the other. My main point is that the model of learning and achievement being used to interpret test results should be consistent with both the design of the assessment procedure and the choice of standard-setting method.

Implications for NAEP

The implications of this analysis for NAEP are a bit ambiguous because NAEP seems to be in a state of transition. NAEP was developed within the context of an analytic model. In its original form, NAEP results were reported item by item, clearly suggesting that the results of individual items could be interpreted in a meaningful way. Furthermore, most of the structure of NAEP still reflects an analytic model. NAEP scaling makes use of Item Response Theory (IRT) models, which assume, for instance, that the information function for the sample of items in a test is equal to the sum of the information functions for the separate items.

Finally, the matrix sampling designs used in NAEP clearly presume that the data can be interpreted within an analytic model; each examinee provides data on a very limited set of tasks, and these data are then combined across examinees and across tasks to provide a relatively complete mosaic. The data collected from each student under NAEP's matrix sampling model is not complete enough to generate dependable score estimates for individuals and is certainly not designed to provide a holistic assessment of achievement for each student. So, I would argue that the operant model for NAEP has been and continues to be an analytic model.

But things are changing, at least in terms of the frameworks and test design. There is increasing reliance on longer, more complex tasks calling for extended responses, and suggesting a holistic model. Of course, the matrix sampling models used for data collection and the IRT models used to estimate score distributions still fit an analytic model much better than they do a holistic model. So the current model is, in most respects, analytic, but has some components that are usually associated with holistic models. Methods may need to be developed that are robust enough to function well in a mixed environment, with some holistic elements and some analytic elements.

PRACTICAL FEASIBILITY

A second criterion on which the two general approaches to standard setting can be compared is the practical feasibility of the methods. That is, how difficult are the judges likely to find the tasks involved? If the tasks required of judges are too difficult for the judges, poor data are likely to emerge no matter how elegant the data collection and analysis procedures are.

Again, there does not seem to be an absolute advantage favoring either the examinee-centered methods or the test-centered methods. Rather, the ease of implementation of the two approaches depends on the context. The degree of difficulty in implementing a standard-setting method is particularly sensitive to the kind of test being used, which is, as indicated earlier, presumably related to the model of achievement that has been adopted. I will consider, as examples, two kinds of tests: objective tests with large numbers of short, discrete items, and performance tests with a few longer and more complex tasks. Again, these two options define the end points of a continuum.

Test-Centered Methods

The test-centered methods, including the Angoff (1971), Nedelsky (1954), Ebel (1972), and Jaeger (1982) methods, were designed for use with objective tests and tend to work fairly smoothly with such tests. It is certainly possible to question whether the results mean what they seem to mean, but the judges apparently do not have great trouble doing what they are asked to do.

We expect to find a fair amount of variability in these ratings because of differences of opinion among the judges on the emphasis given to different topics in instruction and, therefore, on the difficulty of these topics for students, the complexity of the relationship between item characteristics (e.g., the statement of the question or task, the choice of distractors, the placement of questions in a test) and item difficulty, and fluctuations in the judges' attitudes and levels of attention over a long rating process. It is expected that at least some of these sources of variation will be damped by averaging over a large number of tasks or items and over a large number of judges, but they cannot be eliminated.

Nevertheless, the judges asked to complete Angoff ratings managed to do so without complaint in both NAGB standard-setting efforts and in the NAE replication (McLaughlin, 1993b) of one of NAGB studies. The results for NAE's replication were very close to the original NAGB results, and both of these sets of results were in general agreement with the results of NAE's whole-book ratings (McLaughlin, 1993c). Furthermore, the Angoff method has been used to set passing scores on a host of licensure and certification tests, as well as on numerous state testing programs, without major complaints from the judges involved. Not only can judges do the task, but they also do not seem to find it particularly anxiety provoking as such (although they do sometimes find it tedious).

The test-centered methods could be used with performance tests by having the raters review the task and the scoring rubric and then assign a minimum pass level to the task. In some ways, this could be an especially quick and easy way to set the threshold scores, because performance tests typically include only a few tasks. One difficulty that arises in this case is that the rubrics used to score student responses are necessarily fairly general in their specification, and unless the judges are quite familiar with how the rubric was implemented in scoring, they may have trouble estimating in a reasonable way the score to be expected of borderline examinees. As Julian (1993) suggests, the test-centered methods can be used on performance tests, but the use of these methods "fails to capitalize on an advantage of performance examinations, which is the ability of the judges to grasp the task as a whole and to make a holistic judgment about performances" (pp. 3-4). So the test-centered approaches tend to be fairly easy to implement with objective tests, but may be relatively difficult to use effectively with performance tests.

The Examinee-Centered Methods

The examinee-centered methods require judges to evaluate examinee performances. In the case of the contrasting-groups method, the judges assign each examinee to one of two groups, an upper group who have achieved the performance standard and a lower group who have not achieved this standard. In the borderline-group method, the judges identify examinees who just meet the performance standard.

One issue that arises in the use of these methods involves the specific performance to use in assigning examinees to the two groups. As discussed more fully in Technical Note 1 at the end of this paper, either the test performance itself can be used, or some external criterion performance can be used. This choice has potentially important consequences for the interpretation of the test score scale in general, as well as for the threshold score, per se.

In most discussions of examinee-centered methods, it is assumed that the categorization will be based on some general assessment of performance, for example, on ratings by teachers or supervisors, or through a performance test. Where the examinee-centered methods have been used with objective tests, the categorization has usually involved an external criterion that is evaluated by the judges.

Assuming that a few extended performances are used for classification, judges with appropriate experience or training seem quite capable of evaluating each student's performance and assigning the student to one of the groups. The interpretation of the resulting threshold scores might be questionable for various reasons, some of which are discussed later in this paper, but it does not seem that judges have great difficulty in making the categorization decisions required by the examinee-centered methods.

The examinee-centered methods seem much less promising if the categorizations are based on responses to a large numbers of multiple-choice or short-answer questions. Going through long strings of item responses as a basis for categorizing examinees would seem to place severe demands on the memory, attention, and patience of the judges. It may be possible to make this task easier by supplying judges with memory aids like worksheets, but, in general, applying the examinee-centered methods to objective tests seems like a poor use of the judges' time and talent.

So, although either type of standard-setting method can conceivably be used with either kind of test, it seems appropriate to use the examinee-centered method for performance tests (or for a test that is to be interpreted in terms of some criterion performance that can be used as the basis for categorizing examinees). And, it seems appropriate to use the test-centered methods for objective tests. One can pound a nail with a wrench, but it generally does not do much for the nail or the wrench.

Implications for NAEP

As indicated in the last section, the NAEP tests seem to be moving away from an objective format and toward an extended response format. As the tests include more tasks requiring extended responses, it is likely to become somewhat more difficult to use the test-centered methods and somewhat easier to use the examinee-centered methods. This factor is not likely to be decisive, because it is certainly possible to apply the test-centered methods to tests that include both objective items and extended-response items, and several methods have been proposed for adapting the Angoff method to extended-response items. Nevertheless, the examinee-centered methods are likely to become increasingly attractive as NAEP relies more and more on extended-response tasks.

The matrix sampling design employed in NAEP would probably complicate the application of examinee-centered methods and would tend to limit the precision of the threshold scores resulting from the use of examinee-centered methods. Under the matrix sampling design, each examinee responds to a relatively small sample of items, and, therefore, it is likely to be difficult to unambiguously identify examinees as borderline or as being in one of the two contrasting groups. This problem is especially acute because the sample of items in each NAEP booklet is quite limited and is generally not representative of all item types, difficulty levels, and areas of content in the NAEP frameworks.

Matrix sampling does not cause problems for the test-centered approaches because these methods focus on the items, and all of the items in the pool of items for each content area are available for review. In using the test-centered methods, there is no need to evaluate examinee performance.

TECHNICAL ISSUES IN STANDARD SETTING

Both the examinee-centered and test-centered methods have the potential for technical problems. Most of these complications can be controlled with appropriate care and forethought, although some problems are more difficult to control than others.

At least in part because the test-centered methods have been extensively used for more than 20 years, the technical problems associated with these methods are fairly well understood. The evidence seems to suggest that a fairly large number of judges is needed for a precise estimate of the threshold score (Mehrens, 1994; Jaeger, 1991). There is a risk that judges will not have a good sense of how difficult a particular set of items may be for examinees, and, therefore, it has become common practice to provide the judges with data on examinee performance (Jaeger, 1989). There are potential problems with group dynamics (Fitzpatrick, 1989), which users of the Angoff method try to control by structuring the data collection process so that all judges get a chance for active participation.

The NAE report (Shepard et al., 1993) came down hard on the Angoff procedure. In addition to raising a number of technical objections, the NAE panel suggested that the Angoff procedure is fundamentally misguided: "By focussing on items one at a time, for example, the method prevents judges from arriving at an integrated conceptualization of what performance at each of the levels should look like" (p. 72). The technical criticisms were based mainly on five studies conducted by McLaughlin (1993a) and summarized by Shepard et al. (1993) of the internal consistency of the 1992 NAGB ratings.

Consistency of Threshold Scores for Dichotomous and Extended-Response Items

The first of these studies compared the threshold scores set for dichotomously scored items to those set for extended-response items. In the 1992 NAGB studies, the Angoff method was used for the dichotomously scored items, and an examinee-centered method was used for the extended-response items. The resulting threshold scores were highly disparate, with the threshold scores at all three levels (i.e., Basic, Proficient, and Advanced) and all three grade levels being substantially higher for the extended-response items. However, if one accepts the NAE panel's (Shepard et al., 1993) general conclusion that "The weight of evidence suggests that the 1992 achievement levels were set unreasonably high" (p. xxii), it seems more reasonable to reject the results for the extended-response items, which were much higher than those for the Angoff procedure and may have been subject to bias due to a nonrepresentative sample of student papers (see Shepard et al., p. 56), rather than to reject the Angoff results. Given the highly problematic results for the extended-response items, the differences between the Angoff threshold scores and the threshold scores for the extended-response items cannot be counted as evidence against the Angoff procedure.

Consistency in Threshold Scores for Multiple-Choice and Short-Answer Items

The second study involved a comparison between two types of dichotomously scored items, and found a tendency to get higher threshold scores for short-answer items than for multiple-choice items. Shepard et al. (1993) suggest that judges may not have recognized the impact of guessing and the tendency for multiple-choice items to be easier than short-answer items with similar content. Although there are other possible explanations (Kane, 1993), the difficulty in taking differences in item format into account in standard setting probably contributed to the differences between the threshold scores for multiple-choice items and short-answer items. As Shepard et al. (1993) suggest, "When items that appear to be similar in content have very different empirical difficulties, translation into scale values will result in very different cutpoints, unless judges are able to take these format effects into account." (p. 58) However, there is no reason to think that this problem is unique to the Angoff procedure. The judges in examinee-centered standard-setting studies who review examinee responses to items presented in particular formats and contexts are probably not entirely successful in making adjustments for differences in item formats and contexts.

Consistency in Threshold Scores for Hard and Easy Items

The third study compared threshold scores based on ratings of the easiest half of the items to the hardest half of the items, with the effects of multiple-choice versus short-answer format held constant. The threshold scores for the difficult items were consistently higher than the threshold scores for the easy items. McLaughlin (1993a) interpreted these results as evidence that the Angoff method required skills that were beyond the reach of the raters, specifically, "that panelists were not adequately taking into account the differences between easy and hard items, which would seem to be an essential requirement for the validity of the Angoff process" (p. 97).

There are several possible flaws in this line of argument (Kane, 1993). In particular, it assumes that the NAEP scaling system, which is probably the most complicated such system in use, is doing a good enough job in adjusting for item difficulty to ensure that estimates of scores on the theta scale based on the hardest items would generally be equivalent to the estimates based on the easiest items. Research on this kind of vertical equating has not generally been encouraging, and, therefore, it would seem prudent to at least consider the possibility that the scaling procedure might be having an impact on the results of this study. The results of this study (Shepard et al., 1993) may indeed indicate that "panelists tended to underestimate performance on easy items and overestimate it for hard items," (p. 58) but this is not the only possible interpretation.

However, even if the NAE panel's interpretation of the differences between hard and easy items is accepted at face value, it is not clear what this implies for the choice between test-centered and examinee-centered methods. To the extent that judges are unable to take item difficulty into account, all attempts to set absolute standards will be suspect. Judges who use examinee-centered methods to rate the performances of examinees on a set of difficult tasks, but underestimate the difficulty of the tasks, are likely to generate higher threshold scores than judges who rate examinee performances on a set of easier tasks. This problem is especially troublesome in programs like NAEP that employ matrix sampling, because judgments about different examinees may be based on performances on different tasks. As indicated at the end of this paper in Technical Note 1 on criterion effects in examinee-centered standard setting, the need to account for task difficulty is a fundamental issue in examinee-centered standard setting.

So adjustments for task difficulty are an integral, although perhaps implicit, part of any viable standard-setting procedure. Kane and Wilson (1984) suggest that the magnitude of the correlation between item difficulty and the minimum pass levels generated by test-centered methods could be used as one criterion for evaluating these methods. Jaeger (1989) and Kane (1994) suggest that the Angoff method generally does relatively well on this criterion compared to the other test-centered methods. Very little is known about how well the examinee-centered methods can take task difficulty into account.

Consistency Across Content Areas and Cognitive Processes

The fourth study examined how threshold scores varied as a function of content (e.g., the difference between numerical operations and geometry). The NAE panel (Shepard et al., 1993) expected threshold scores "based on unfamiliar subjects to be set higher in relation to the empirical scale than those based on the familiar content strand" (p. 60). The fifth study examined differences in threshold scores associated with different cognitive processes (e.g., procedural knowledge versus problem solving). The NAE panel expected differences in threshold scores for this dimension because they expected that the judges would wish to see greater gains in the future on items demanding higher levels of reasoning and would therefore set higher standards for these items.

The differences expected by the NAE panel (Shepard et al., 1993) in the fourth and fifth studies were not found, and the panel interpreted this as further evidence that judges cannot adequately perform the tasks required of them by the Angoff procedure. However, it could also be argued that the judges imposed what they considered to be reasonable expectations for the different content areas and different cognitive processes. That these expectations were not consistent with the panel's expectations does not necessarily imply that the judgments were faulty.

Implications for Choice of Standard-Setting Method

The general model employed by McLaughlin (1993a) and by Shepard et al. (1993) seems to view differences in results between dichotomous items and extended-response items, between multiple-choice and short-answer items, and between easy and hard items to be method artifacts that should have been eliminated by the judges, and to view differences associated with content and cognitive process as real differences that should have been reflected in the ratings.

This is not an unreasonable point of view, but it is not the only reasonable point of view. One could assume that the apparent differences in the quality of student performance between dichotomous items and extended response items and between multiple-choice items and short-answer items are real, and that students are meeting the judges' expectations on recognition tasks, but are not doing as well, relative to the judges' expectations, on tasks that require an extended response. If it is a given that multiple-choice items will always yield the same results as supply-type items, it is not clear why NAEP would use anything but multiple-choice items, which are quick, easy to score, relatively easy to scale, and relatively inexpensive. The fact is that many scholars believe extended-response items tap aspects of student achievement not directly assessed by multiple-choice items.

The results of the five studies reported by McLaughlin (1993a) do point to some possible problems in test-centered methods. If it can be shown (by eliminating other possible explanations) that the differences between the threshold scores for hard and easy items are actually due to a tendency on the part of judges to avoid extreme values in implementing test-centered methods, a potential source of systematic error in the resulting threshold scores would be indicated. The discrepancies between results for dichotomous items and extended response items may pose a practical problem whenever both kinds of items are used on the same assessment. It is not clear that the other three studies point to any serious limitations in the Angoff procedures. The evidence developed in the five studies of the technical properties of the 1992 NAEP standard setting do not seem to justify the conclusion (Shepard et al., 1993) based largely on these studies "that the Angoff procedure is fundamentally flawed because it depends on cognitive judgments that are virtually impossible to make" (p. 77).

Furthermore, one NAE study provided support for the generalizability of the threshold scores obtained with the Angoff method. McLaughlin (1993b) replicated the Angoff process for 50 items used in the eighth-grade mathematics assessment in 1992, with some changes in the order in which the threshold scores were set. The fact that the Angoff ratings obtained in McLaughlin's replication yielded results that were very similar to those obtained in the original NAGB study tends to support the generalizability of the Angoff ratings over time, samples of raters, and specific implementation procedures. The fact that a new group of panelists, using a somewhat different methodology, and under different leadership, came up with roughly the same threshold scores as NAGB mathematics panel, provides very strong support for the connection between the achievement-level descriptions and the corresponding threshold scores.

Cronbach (1971) suggested that correspondence of a test to a test plan could be evaluated by having two test-development teams independently develop tests from the same test plan, and then comparing examinees' scores on the two tests. In this case, the two independent Angoff standard-setting studies have been found to set essentially the same threshold scores. Few, if any, test-development efforts have been subjected to this kind of rigorous evaluation in the last 23 years.

Technical Issues for Examinee-Centered Methods

The technical problems that are likely to occur in the context of examinee-centered methods have not been fully explored because these methods have not been used much in high-stakes settings. The sample sizes needed, both for judges and examinee papers, in order to get dependable results using the examinee methods have not been determined. There are issues to be addressed in the sampling of papers. (e.g., Is it better to use a random sample, a sample with a uniform distribution of papers over the score range, or some other sampling plan?) What kind of feedback, if any, should be given to judges about the consequences of setting a threshold score at a particular point? If the judges discuss their ratings at any point, all of the complexities of group interactions become relevant.

The two technical notes at the end of this paper present analyses of two specific technical problems that apply to the examinee-centered methods. The first problem is associated with the choice of a criterion for categorizing examinees as borderline or as being in the upper or lower contrasting groups. The analyses in Technical Note 1 indicate that the threshold score resulting from an examinee-centered standard-setting study will depend on the choice of criterion. To the extent that the criterion used to categorize examinees is inconsistent with the intended interpretation of the test score, the decisions based on the threshold scores will be inconsistent with this interpretation. If the criterion is consistent with the intended interpretation, it seems reasonable to expect that the test scores be validated against this criterion. In general, the analyses in the first note suggest that the selection of criteria for examinee-based standard setting deserves more attention than it has typically gotten in the past.

Technical Note 2 at the end of this paper examines the potential impact of student misclassification in contrasting-groups studies. The analysis indicates that random errors leading to misclassifications introduce a bias into the resulting threshold scores. In particular, these misclassifications lead to a regression of all threshold scores toward the mean of the distribution of student scores.

As more research is done on these methods, other potential problems are likely to be identified. No doubt, ways to solve most of these problems, or at least ameliorate their consequences, will be found, but it would clearly be a triumph of optimism over experience to assume that the examinee-centered methods will not encounter problems.

Conclusions

There are at least three levels on which examinee-centered and task-centered standard-setting methods can be compared: (a) in terms of consistency with the model of achievement underlying test design and interpretation in a specific context, (b) in terms of practical feasibility in the context, and (c) in terms of the technical considerations that apply to the context.

The main conclusion to be drawn from the analyses in this paper is that neither the examinee-centered nor the test-centered method is best in all contexts. Each type of standard-setting method has strengths and weaknesses that make it appropriate in some situations and inappropriate in others.

The two kinds of standard-setting methods are associated with fundamentally different models of achievement and the assessment methods usually associated with these achievement models. The test-centered methods seem to be particularly appropriate for use with tests consisting of multiple, independently-scored tasks within the context of an analytic model of achievement. Examinee responses to individual tasks are scored separately, and these item scores are combined to yield the examinee's overall score. The test-centered methods parallel this scoring procedure by assigning a minimum pass level to each task and then combining these item MPLs into an overall threshold score. The examinee-centered methods provide a particularly good fit for performance tests within the context of a holistic model. Matching the standard-setting method to the model of achievement and learning would seem to be a fundamental concern if the overall interpretation is to be coherent.

The two kinds of standard-setting methods also differ in how easily they can be applied to different kinds of tests. Judges are likely to find it difficult to apply the examinee-centered methods to objective tests of any length but have no trouble applying the test-centered methods in this context. Conversely, judges may find it difficult to apply the test-centered methods to extended-response performance tests without having experience in using the scoring rubrics but have no difficulty in applying the examinee-centered methods in this context.

All standard-setting methods have potential technical problems. Currently, the examinee-centered methods seem to have more problems that are not fully explored, simply because these methods have not been examined as closely as the test-centered methods, but this imbalance could be corrected fairly quickly.

TECHNICAL NOTE 1

CRITERION EFFECTS IN EXAMINEE-CENTERED STANDARD SETTING

The contrasting-groups method requires judges to categorize examinees based on judgments of their knowledge and skills without using the test scores (Livingston & Zieky, 1982), but, except for the stipulation that the test scores per se not be used, the nature of the criterion is quite open. There is nothing in the logic of the method that precludes the use of judgments about examinee performance on the test to assign examinees to the two contrasting groups. However, most writers (e.g., Jaeger, 1989; Shepard et al., 1993) have described the examinee-centered (or "judgmental") methods in terms of judgments about examinee performance outside of the testing context. And most applications have been based on performance assessments or some kind of rating of examinee performance (e.g., ratings by teachers). The study by McLaughlin et al. (1993) introduced two possible criteria, one based on teacher evaluations of their own students with reference to NAGB achievement level descriptions and one based on one-on-one assessments of a sample of students by the researchers conducting the study. There are many possibilities. And, as noted earlier in this paper, the examinee-centered methods tend to be most easily implementable for performance tests with a limited number of tasks.

However, it is possible to use the test performance itself (even if the test is an objective test) as the criterion (Clauser & Clyman, 1994; McLaughlin, 1993c). As indicated later in this note, if the test scores are to be interpreted in terms of test performance (as NAEP usually is), it is desirable to use the test performance as the criterion.

One way to explore the consequences of choosing various criterion measures is to examine the likely results of a series of "thought experiments" in which different choices are made. Thought experiments, or "gedanken experiments" (Kaplan, 1964; Hempel, 1965), have a long and distinguished history, stretching back at least to Galileo. The basic approach is to explore the implications of a theory or model by imagining what must happen, or is likely to happen, under certain circumstances. The purpose is to clarify the meaning of constructs and models by exploring their implications in special cases and to check for inconsistencies or ambiguities in the models and constructs. In order to derive clear, specific conclusions from the thought experiments, the experiments must usually be idealized by ignoring many factors that would have to be controlled in real experiments, and in many cases, by considering limiting cases (e.g., frictionless surfaces, isolated systems) that can be approximated but not realized in actual experiments.

Thought experiments provide a convenient way to analyze the impact of the criteria used to classify examinees into the borderline group or into the two contrasting groups. In developing these thought experiments, I will make use of several general assumptions about how an experiment *might* turn out under the assumptions specified. I am not assuming that these conditions always hold, or even usually hold. All that is necessary for the thought experiment to be useful is that the conditions could occur in some cases.

The first assumption is that the performance standard, or achievement level description, is likely to be stated in fairly general terms, without reference to the specifics of data collection. The descriptions are likely to focus on types of activities that passing examinees can perform and that failing examinees cannot perform, without specifying all of the conditions of observation. For example, the item format (multiple-choice, short-answer, interview), the context (classroom, testing center), the conditions of measurement (time limits, access to reference materials), and the examiner (teacher, parent, external examiner) are not likely to be included in the descriptions of the performance standards.

Second, the kinds of activities that examinees can perform will depend in part on the context and the data collection methods. The child who makes sense of a story while reading with a parent or teacher in a familiar classroom setting may not be able to do nearly as well during a standardized testing session in an auditorium. The examinees who can solve an area problem presented in a standard format that has been explicitly taught and practiced may not be able to solve a more "authentic" area problem, even if the more authentic problem could, in principle, be solved using the same arithmetic operations as the standard problem. Conversely, a child who can solve a problem formulated in familiar terms (e.g., making change) might have trouble with an abstract version of the "same" problem.

Third, assuming that Y and Y' represent two possible criterion measures involving different contexts and conditions of observation, with different judges evaluating performance, it should not be surprising that systematic differences in performance on these two criteria are found. That is, examinees might generally do better on Y' than on Y (in the sense that examinees tend to achieve greater success on various tasks in the context of Y' than in the context of Y), or examinees might do better on Y than on Y'. For the sake of definiteness, I will assume that performance tends to be better on Y' than on Y. That is, Y' is an easier or perhaps more supportive mode or context of assessment than Y. If Y' is substantially easier than Y, it would be reasonable to expect that there may be a substantial number of examinees who would satisfy the performance standard if evaluated using Y' but would not satisfy the performance standard if evaluated using Y, and that there would be few if any examinees who would satisfy the standard under Y but would not satisfy it under Y'.

Fourth, for each of these criterion measures, I will assume that the probability that an examinee satisfies the requirements in the performance standard is an increasing function of the examinee's score on the score scale on which the threshold score is to be defined. Without this assumption, the use of the contrasting-groups method does not make sense. Further, it is convenient to assume that for Y and Y', the probability that examinees satisfy the performance standard is relatively low toward the bottom of the score scale and relatively high at the high end of the score scale, and that it increases fairly rapidly in some interval on the score scale. (An increasing logistic function like those used to represent item response functions satisfy these requirements, but it is not necessary to pick any particular model.)

Now, I will consider three points on the score scale, A, B, and C, such that:

$$A < B < C.$$

I will assume that students with scores around C are classified as having satisfied the performance standard embodied in the achievement level description on both Y and Y'. The students with scores around point C satisfy the performance standard under the supportive conditions, Y', and the less supportive conditions, Y. These students can perform the kinds of activities in the achievement level description even in a relatively unsupportive context. At the other extreme, students with scores around A do not meet the standard even in the relatively supportive context, represented by Y'.

Students with scores around B are generally judged to have satisfied the performance standard specified in the achievement level descriptions, if assessed on Y', but are generally judged as having not satisfied the standard if assessed under Y. Students with scores around B are, in a sense, in a transitional stage. They are able to meet the performance standard in a supportive context like Y', but are not able to meet this standard in an unsupportive context like *Y*. Given the assumption that scores on *Y′* are generally higher than those on *Y*, it seems reasonable to assume that there will be very few students who meet the requirements under *Y* but not under *Y′*.

The claim being made here is not that these assumptions hold in all cases where one might want to employ the contrasting-groups method, nor that they hold in any particular case. Rather, all that is necessary for the thought experiment described below to clarify the implications of the choice of criterion measure, is that these assumptions could apply to some applications of the contrasting-groups method. It is, for example, not necessary to actually develop a supportive criterion assessment and an unsupportive criterion; it is only necessary to accept the possibility of developing two criteria, one of which is more difficult than the other.

I will make three points, by way of the thought experiment based on the assumptions outlined above. The first point is that the choice of the criterion used to classify examinees in the examinee-centered standard-setting methods can potentially make a difference in the resulting threshold score. The second point is that the choice of the criterion needs to be tied to the intended interpretation of the test scores in order to avoid inconsistencies in the interpretation of the results. The third point is that the intended interpretation should presumably be supported by appropriate validation evidence, including, perhaps, evidence from the standard-setting study.

Impact of the Criterion on the Threshold Score

In the thought experiment, the borderline-group method and the contrasting-groups method are used to set a passing score. Consider first what is likely to happen if the *Y* assessment is used to categorize examinees in the standard-setting study. Under the conditions included in *Y*, the examinees with scores around score points *A* and *B* tend not to achieve the performance standard specified in the achievement level descriptions, and the examinees with scores around *C* tend to achieve this standard. Therefore, in the borderline-group study, most of the examinees identified as borderline examinees are likely to have scores between *B* and *C*, and the passing score is likely to be between *B* and *C*. Under the contrasting-groups method, most of the examinees in the upper group are likely to have scores around or above point *C*, and most of the lower group will tend to have scores around or below point *B*, and, therefore, the passing score would again be expected to be set somewhere between *B* and *C*.

Now, consider what is likely to happen if *Y′* is used to categorize examinees. Under *Y′*, most examinees at point *A* fail to satisfy the performance standard, while most examinees at points *B* and *C* meet this standard. The borderline examinees are likely to have scores between *A* and *B*, and, therefore, the threshold score is likely to be between *A* and *B* for a borderline-group study. Under the contrasting-groups method, it is reasonable to assume that most of the examinees in the lower group will have scores around or below *A* and most of the examinees in the upper group will have scores around or above point *B*, again yielding a passing score between *A* and *B*.

The net result is that if the easier criterion, *Y′*, is used to classify examinees into contrasting groups or into the borderline group, the passing score tends to be lower, and, therefore, more examinees pass than if the harder criterion, *Y*, were used. In particular, given the assumptions built into the thought experiment, if *Y′* is used, the passing score is between *A* and *B*, and, therefore, the examinees with scores around *B* pass. If *Y* is used, the passing score is between *B* and *C*, and, therefore, the examinees with scores around *B* fail. In general, the choice of the criterion for categorizing examinees can have a strong impact on the threshold score, and, therefore, in order to avoid arbitrariness, some basis for choosing the criterion measure is needed.

The Choice of Criterion

The next question involves a choice between different possible criteria, say between *Y* and *Y′*. Given that these two criteria yield different passing scores, which criterion should be preferred? The answer is contingent on the interpretation of the test scores.

If scores are interpreted in terms of performances like those included in *Y*, examinees with scores around *B* have not satisfied the performance standard, because the *Y* assessments indicate that most of these examinees cannot perform the kinds of activities included in the performance standard. That is, direct observation of examinee performance on the criterion associated with the intended interpretation of the test scores indicates that examinees with scores around *B* should generally not pass. If *Y* is used to set the passing score, the passing score would probably be above *B*, and, therefore, examinees with scores around *B* would not pass and, therefore, do not meet the standard. The decisions based on direct observation of the criterion, *Y*, are in agreement with decisions based on test scores (because the test-based decision rule is derived from the same criterion).

However, if *Y′* is used as the criterion in an examinee-centered standard-setting study, the passing score would be set below *B*, and, therefore, examinees with scores around *B* will pass. This result would be inconsistent with decisions based directly on the criterion, *Y*, associated with the test score interpretation.

Similarly, if scores are interpreted in terms of *Y′*, and *Y′* is used to classify examinees in a contrasting-groups or borderline-group study, the results will be consistent. However, if scores are interpreted in terms of *Y′* and the passing score is set using *Y* as the criterion, the results will be inconsistent. In general, in order to avoid inconsistency in the interpretation of the results, the criterion used to assign examinees to categories for the examinee-centered methods should be consistent with the intended interpretation of the test scores.

Implications for Validity

It follows from this conclusion and the requirements in the Standards for Educational and Psychological Tests (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985), that the criterion used in the examinee-centered standard-setting study should be consistent with the test score interpretation and, therefore, that the test scores should be validated in terms of this criterion. That is, if it is appropriate to interpret test scores in terms of performance on the criterion assessment used in a standard-setting study, then the results should be reported in terms of this criterion, and the test scores should be validated in terms of this interpretation, perhaps employing a criterion-validation study (or perhaps using some indirect, construct-based validation strategy).

This analysis indicates that considerably more care is needed in choosing criterion assessments for examinee-centered standard-setting studies than has generally been recognized. The analysis also indicates the importance of doing the research necessary to validate the test scores against the criterion used in standard setting.

This kind of validation research requires additional effort, but the effort is likely to be worthwhile. A careful analysis of the criterion to be used for standard setting could have the important positive consequence of encouraging greater clarity in the proposed interpretation of the test scores. And if it is feasible to use the criterion to set the passing score, it is presumably feasible to use this criterion in a validation study. Given the dearth of validity evidence for most testing programs, the possible availability of defensible criterion-related validity evidence presents an opportunity that is not to be missed. The criterion is likely to be considered a partial criterion, one that serves as a good indicator of the achievement of interest rather than as a definition of this achievement, but this kind of convergent evidence is potentially very valuable. (Note that if the criterion measure does not reflect the intended interpretation well enough to be useful as a potential criterion measure in validation studies, its appropriateness as a criterion for standard setting should also be questionable.)

So, the criterion effect does not necessarily cast doubt on the validity of the passing score. Rather, a careful consideration of this effect may strengthen the validity evidence for the intended interpretation of the test scores. However, there is a price to be paid for this benefit. The criterion used in the examinee-centered standard-setting study needs to be consistent with the intended interpretation, and the test scores should be empirically validated against this criterion, if such validation is feasible.

Implications for NAEP

The implications for NAEP of the issues discussed above are potentially important. If the intended interpretation of NAEP results is in terms of student performance on the NAEP tests, then one of the main studies underlying the conclusions in the NAE study is subject to a criterion bias of unknown magnitude and direction, because the two criteria used in that study were both quite different from the NAEP tests.

If on the other hand, the intended interpretation of NAEP scores is in terms of teacher evaluations of student performance or in terms of performance on special one-on-one criterion assessments, then it would seem reasonable to expect that the NAEP scores be validated for this interpretation, and that NAEP data be reported in terms of predictions about the kinds of performance included in these criterion variables. Although the amount of data is limited, it appears from the data in Tables 4.3 and 4.4 on page 86 of the NAE report (Shepard et al., 1993), that examinees generally did much better on the alternative criterion measures used in the NAE study than they did on the NAEP tests. If this is true, and if the intended interpretation of NAEP results is in terms of criteria like those in the NAE study, NAEP results (whether reported in terms of achievement levels, anchor-point descriptions, or exemplar items) would be systematically underestimating examinee achievement.

There is a question of some importance here. Traditionally, the reporting of NAEP results seems to have been in terms of performance on the test items; initially results were even reported item by item. However, a good case can be made for having the reported NAEP results reflect what students can do in a variety of nontest contexts (e.g., classrooms or community settings). However, if NAEP results are to be interpreted in this way, we probably need to consider either changing the nature of NAEP data collection to incorporate these criterion performances as part of the NAEP and/or conducting the kind of criterion-related studies that will support inferences from the current NAEP tests to teacher judgments and the results obtained on one-on-one performance tests.

TECHNICAL NOTE 2

SELECTION EFFECTS IN EXAMINEE-CENTERED STANDARD SETTING

The selection of specific groups of examinees to represent the different categories in examinee-centered studies may also have an impact on the results. This effect is likely to have a much larger impact on the results of a contrasting-groups study than on a borderline-group study, and, therefore, the discussion in this note will be of the contrasting-groups method.

Suppose that we have a group of students for whom we have scores on some score scale, and we want to use the contrasting-groups method to set a threshold score on this scale. There are several ways in which this could be done. We could have the judges classify students into the upper contrasting group or into the lower contrasting group depending on whether or not they are judged to have met the performance standard. The passing score can then be set at the scale point such that the number of misclassifications is the same in both directions. This certainly seems sensible and corresponds, in general terms, to the procedure used in the contrasting-groups study conducted by McLaughlin et al. (1993).

Let's assume that we are setting a single passing score on a test and that the observed score distribution is approximately normal. Assume further that the performance standard is such that most examinees are expected to pass; that is, the threshold score is at the low end of the score distribution. It is also convenient to assume that the test scores are perfectly reliable, in order to avoid complications that would result from an explicit consideration of errors of measurement in the test scores.

We are to use the contrasting-groups method. We will assume that the judges use some criterion that is perfectly correlated with the test score, except for random errors that sometimes result in misclassification, that is, there are no systematic errors in the criteria being used by the judges.

So, there is some point on the score scale, call it C, such that examinees with scores above C have met the criteria in the performance standard and, therefore, should be categorized as passing, and examinees with scores below C have not met the criteria in the performance standard and, therefore, should be categorized as failing. C marks the true boundary between those who have met the standard and those who have not met the standard.

We will allow for the possibility that the judges may misclassify examinees because of random errors in their assessments. The point of this thought experiment is to examine the impact of such misclassifications on the resulting threshold score. However, we will assume that the judges are not biased in either direction; that is, they are as likely to misclassify an examinee with a score some distance below the threshold score as they are to misclassify an examinee with a score that is the same distance above the threshold score.

Let x represent the distance from the cut score to some test score (as a distance, x is greater than or equal to zero). Let $f_{1,x}$ be the frequency of scores in some interval, dx, that is x units below the point,

C, for examinees in the lower contrasting group. So, the total number of individuals in the lower group is:

$$N_1 = \int_0^\infty f_{1,x} dx. \tag{1}$$

Similarly, we can let $f_{2,x}$ be the frequency for the distribution of observed scores in the upper contrasting group. So, the total number of individuals in the upper group is:

$$N_2 = \int_0^\infty f_{2,x} dx. \tag{2}$$

N_1 is the number of examinees who are actually in the lower group and should be classified in the lower group. N_2 is the number of examinees who are actually in the upper group and should be classified in the upper group.

However, we are allowing for the possibility of misclassification due to unreliability in the judges' decisions. Let p_x represent the probability that an examinee with a score at a distance of x from the threshold score is classified correctly, and let q_x represent the probability that this examinee is classified incorrectly. For each value of x,

$$p_x + q_x = 1.0. \tag{3}$$

For large values of x, the probability of misclassification will presumably be small, and the probability of correct classification will approach 1.0. Note that these probabilities are associated with distance from the threshold score, not direction. We are assuming that the situation is symmetric in the sense that the probability of misclassification is the same for examinees x units below C as it is for examinees who are x units above C.

Because the distribution of scores is approximately normal, and the threshold score, C, is below the mean, we can assume that, for any value of x:

$$f_{1,x} \leq f_{2,x}. \tag{4}$$

For small values of x, we have scores that are close to each other, on either side of C, and, therefore, $f_{2,x}$ will be only slightly larger than $f_{1,x}$. For larger values of x, $f_{2,x}$ will be much larger than $f_{1,x}$.

The number of examinees classified into the lower group by the judges will be equal to the number who are actually in the lower group and are classified correctly, plus the number in the upper group who are misclassified:

$$N'_1 = \int_0^\infty f_{1,x} p_x dx + \int_0^\infty f_{2,x} q_x dx. \tag{5}$$

Because $f_{2,x}$ is greater than $f_{1,x}$ for all values of x, except $x = O$, where they are equal, we can write:

$$\begin{aligned} N'_1 &> \int_o^\infty f_{1,x} p_x dx + \int_o^\infty f_{1,x} q_x dx \\ &= \int_o^\infty f_{1,x} (p_x + q_x)\, dx \\ &= \int_o^\infty f_{1,x} dx \\ &= N_1. \end{aligned} \tag{6}$$

That is, under the conditions specified, the number of examinees classified into the lower category will be greater than the number that should be classified there. Since the total number of examinees has not changed, we also know that the number of examinees classified into the upper category will be lower than the number that should be classified there. The fact that the number of examinees in the upper category will be underestimated can also be shown directly in a manner analogous to that used above to show that the number in the lower category will be overestimated.

Assuming that the threshold score is chosen so that the number of false-positive errors equals the number of false-negative errors, the threshold score will tend to be set higher than it should be. The magnitude of this bias toward higher threshold scores is a function of the overestimation of the number of individuals in the lower category, which is given by:

$$N'_1 - N_1 = \int_o^\infty (f_{2,x} - f_{1,x}) q_x dx. \tag{7}$$

If q_x is zero for all values of x, no classification errors are made, and the obtained threshold score equals the actual threshold score, C. As the probability of misclassification increases for various values of x, the magnitude of the difference between N_1 and N'_1 increases, and the bias in the threshold score increases. In the limiting case where the judges are operating at the chance level, p_x and q_x would both be equal to 0.5 for all values of x, and in order to have equal numbers of false positives and false negatives, the threshold score would be set at the mean. In effect, the estimates of those threshold scores that are below the mean are regressed toward the mean.

This effect is, of course, not restricted to the low end of the distribution. Threshold scores that are above the mean will be underestimated. In all cases, the classification errors associated with unreliable judgments will tend to regress the estimated threshold scores toward the mean of the distribution.

The assumptions employed in this argument are needed so that the mathematical derivation will go smoothly. The underlying mechanism is quite simple (i.e., there are more false-negative errors than false-positive errors because there are more people in the upper group who can be misclassified), and, therefore, the bias toward less extreme threshold scores is likely to be present whenever the threshold score is not at the mean of the distribution of scores.

Implications for NAEP

The analyses in this technical note and in Technical Note 1 have identified two technical problems in the examinee-centered methods. As the examinee-centered methods are subjected to closer scrutiny, other problems are likely to be identified. Although the examinee-centered methods have been around for a long time, they have not been widely used in high-stakes testing programs, and therefore their properties have not been examined much by researchers. It may be the case that most of these problems can be addressed by either modifying procedures or by using statistical adjustments, but progress in solving these problems will not occur until the various sources of bias in these procedures are recognized.

At least for the present, substitution of one of the examinee-centered methods for the Angoff procedure will not eliminate questions about the technical properties of the resulting threshold scores. All of the standard-setting methods in current use have potential problems. The Angoff procedure enjoys the advantage of having had its properties examined in a relatively large number of research studies (See Jaeger, 1989; Mehrens, 1994), and to have fared relatively well across this series of studies.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement (2nd ed.*, pp. 508-600). Washington, DC: American Council on Education.

Cizek, G. J. (1993). *Reactions to National Academy of Education report, Setting performance standards for student achievement.* Washington, DC: National Assessment Governing Board.

Clauser, B., & Clyman, S. (1994). A contrasting groups approach to standard setting for performance assessments of clinical skills. *Academic Medicine* [RIME Supplement].

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed, pp. 443-507). Washington, DC: American Council on Education.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438-481.

Ebel, R. L. (1972). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Fitzpatrick, A. R. (1989). Social influences in standard setting: The effects of social interaction on group judgments. *Review of Educational Research, 59,* 315-328.

Hempel, C. G. (1965). *Aspects of scientific explanation*. New York: Free Press.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4,* 461-475.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 485-514). New York: American Council on Education/Macmillan.

Jaeger, R. M. (1991). Selection of judges for standard-setting. *Educational Measurement C. U. Issues and Practice, 10,* 3-6, 10, 14.

Julian, E. R. (1993, April). *Standard setting on performance examinations*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Kane, M. (1993). *Comments on the NAE evaluation of NAGB achievement levels.* Washington, DC: National Assessment Governing Board.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64,* 425-461.

Kane, M., & Wilson, J. (1984). Errors of measurement and standard setting in mastery testing. *Applied Psychological Measurement, 8,* 107-115.

Kaplan, A. (1964). *The conduct of inquiry: Methodology for behavioral science*. New York: Harper and Row.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

McLaughlin, D. H. (1993a). Validity of the 1992 NAEP achievement-level-setting process. In *Setting performance standards for student achievement: Background studies*. Stanford, CA: Stanford University, National Academy of Education.

McLaughlin, D. H. (1993b). Order of Angoff ratings in multiple simultaneous standards. In *Setting performance standards for student achievement: Background studies*. Stanford, CA: Stanford University, National Academy of Education.

McLaughlin, D. H. (1993c). Rated achievement levels of completed NAEP mathematics booklets. In *Setting performance standards for student achievement: Background studies*. Stanford, CA: Stanford University, National Academy of Education.

McLaughlin, D. H., DuBois, P., Eaton, M., Ehrlich, D., Stancavage, F. B., O'Donnell, C., Yu, J., and DeStefano, L. (1993). Comparison of teachers' and researchers' ratings for students' performance in mathematics and reading with NAEP measurement of achievement levels. In *Setting performance standards for student achievement: Background studies*. Stanford, CA: Stanford University, National Academy of Education.

Mehrens, W. A. (1995). Methodological issues in standard setting for educational exams. In *Proceedings of Joint Conference on Standard Setting for Large-Scale Assessment* (pp. 221-263). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Mullins, M., & Green, D. (1994, Winter). In search of truth and the perfect standard-setting method: Is the Angoff procedure the best available for credentialing. *Clear Exam Review*, 21-24.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement*. National Academy of Education, Stanford, CA: Stanford University.

Wiggens, G. P. (1993, November). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 75,*(2), 200-214.

Implications for Standard Setting of the National Academy of Education Evaluation of the National Assessment of Educational Progress Achievement Levels

Lorrie A. Shepard

Professor of Education, University of Colorado at Boulder

ABSTRACT

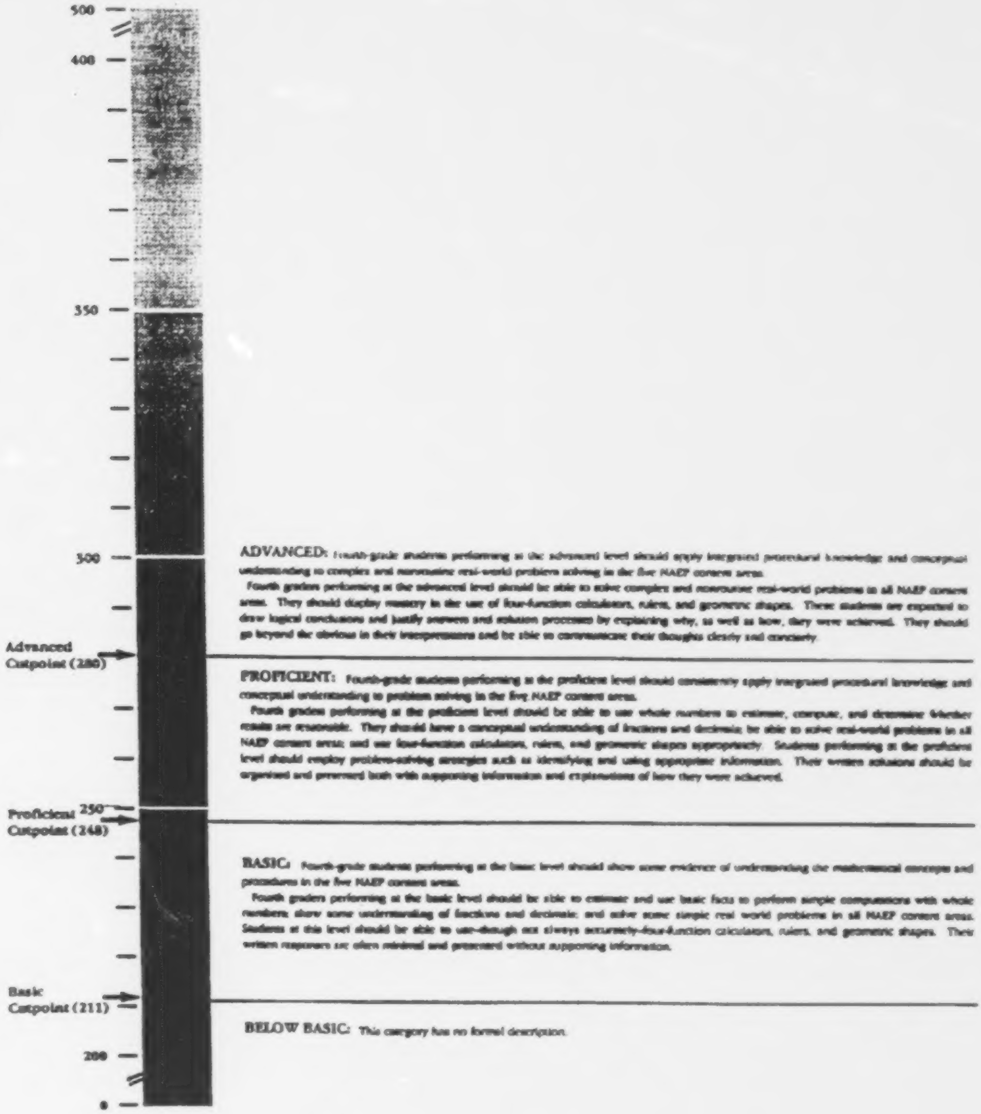
A summary is provided for key findings from the National Academy of Education (NAE) Panel's evaluation of the National Assessment of Educational Progress (NAEP) achievement levels. In the 1992 assessments, items used to exemplify performance failed on both substantive and statistical criteria. External validity studies and internal reanalyses demonstrated inadequacies in the item judgment method used to translate judges' intended standards onto the score scale. Therefore, NAEP achievement level results did not accurately report what students can or cannot do. In the future, assessment tasks should be developed directly to represent intended performance levels; and more holistic standard-setting approaches should be used to allow judges to focus on substantive expectations and avoid false assumptions about item intercorrelations and item difficulties.

BACKGROUND

Achievement levels set for NAEP are examples of a type of "performance standard" used to report and interpret assessment results. In 1992, achievement levels were set for NAEP in both reading and mathematics. Within each grade level, three achievement levels were specified: advanced, proficient, and basic. By implication, a fourth score category was created which signified below-basic performance. As illustrated in Figure 1 for fourth-grade mathematics, each achievement level was accompanied by a verbal summary describing performance at that level. To set the achievement-level standards, representative panels of educator and noneducator expert judges first developed the narrative descriptions. Then, judges followed a technical procedure known as the modified-Angoff method to select the specific cut scores separating the achievement levels. For example, fourth graders earning a score of 280 or above are considered advanced.

This paper provides a summary of key findings from the National Academy of Education (NAE) evaluation of the 1992 NAEP achievement levels. The NAE Panel (the Panel) was originally created in 1990 in response to a mandate from Congress for an independent evaluation of the effects of extending NAEP to the state level. Because of the salience of standards in educational reform and controversy surrounding earlier evaluations of the NAEP achievement levels, the National Center for Education Statistics (NCES) asked the NAE Panel to expand its work and conduct an evaluation of the 1992 achievement levels for reading and mathematics.

BEST COPY AVAILABLE



ADVANCED: Fourth-grade students performing at the advanced level should apply integrated procedural knowledge and conceptual understanding to complex and nonroutine real-world problem solving in the five NAEP content areas.

Fourth graders performing at the advanced level should be able to solve complex and nonroutine real-world problems in all NAEP content areas. They should display mastery in the use of four-function calculators, rulers, and geometric shapes. These students are expected to draw logical conclusions and justify answers and solution processes by explaining why, as well as how, they were achieved. They should go beyond the obvious in their interpretations and be able to communicate their thoughts clearly and concisely.

PROFICIENT: Fourth-grade students performing at the proficient level should consistently apply integrated procedural knowledge and conceptual understanding to problem solving in the five NAEP content areas.

Fourth graders performing at the proficient level should be able to use whole numbers to estimate, compute, and determine whether results are reasonable. They should have a conceptual understanding of fractions and decimals; be able to solve real-world problems in all NAEP content areas; and use four-function calculators, rulers, and geometric shapes appropriately. Students performing at the proficient level should employ problem-solving strategies such as identifying and using appropriate information. Their written solutions should be organized and presented both with supporting information and explanations of how they were achieved.

BASIC: Fourth-grade students performing at the basic level should show some evidence of understanding the mathematical concepts and procedures in the five NAEP content areas.

Fourth graders performing at the basic level should be able to estimate and use basic facts to perform simple computations with whole numbers; show some understanding of fractions and decimals; and solve some simple real world problems in all NAEP content areas. Students at this level should be able to use—though not always accurately—four-function calculators, rulers, and geometric shapes. Their written responses are often minimal and presented without supporting information.

BELOW BASIC: This category has no formal description.

Figure 1. NAEP fourth-grade math achievement level descriptions and cut points.

Note. Reprinted from *Setting Performance Standards for Student Achievement: A Report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: An Evaluation of the 1992 Achievement Levels* (p.34), by L. Shepard, R. Glaser, R. Linn, and G. Bohrnstedt, 1993, Stanford, CA: National Academy of Education.

BEST COPY AVAILABLE

The Panel's studies were extensive, including more than a dozen separate field studies and reanalysis of existing data. Findings from the Panel's report, Setting Performance Standards for Student Achievement (Shepard, Glaser, Linn, & Bohrnstedt, 1993), have important implications for future efforts to set achievement levels on NAEP and for efforts to set performance standards for large-scale assessments more generally. For purposes of this paper, I focus on two key issues: (a) the adequacy of sample items and descriptions used to exemplify performance for each of the achievement levels, and (b) the adequacy of the Angoff procedure for translating judges' expectations into cut scores. Although both of these problem areas were the source of serious deficiencies in the 1992 implementation, the first problem has a relatively straightforward solution. The second problem is more controversial and does not have a guaranteed solution. Nonetheless, some conceptual guidelines can be inferred from the Panel's studies. In the third and final section of the paper, I discuss models of expertise and the kinds of procedures needed to involve judges directly in setting quantitative standards that correspond to their substantive expectations.

THE IMPORTANCE OF EXEMPLARY SAMPLE ITEMS AND SUBSTANTIVE DESCRIPTIONS

In an era of standards-based educational reform, there are multiple meanings for the term standard. Content standards, exemplified by the National Council of Teachers of Mathematics (NCTM) Standards in mathematics, are like curriculum frameworks. They specify what should be taught at each level of schooling. Performance standards are more specific and designate what students must do to demonstrate minimum, satisfactory, or superior levels of performance. Although other forms of performance standards are possible, achievement levels set for NAEP are one type of performance standard. Expressed as a cut score on the total-score scale, the achievement levels specify what a student must be able to do to be considered basic, proficient, or advanced.

On a large-scale assessment such as NAEP, performance standards are not used to report individual scores or to certify student mastery. Rather, NAEP is intended to monitor national trends in achievement. In this context, the National Assessment Governing Board's (NAGB) purpose in setting achievement levels was to make assessment results more meaningful and, therefore, more useful in advancing educational reform. According to Chester Finn (1989), NAGB Chairman in 1989, "NAEP has long had the potential not only to be descriptive but to say how good is good enough. . . . In the spirit of Charlottesville the Board has now started moving to establish benchmarks for learning, to use with its tests." (p. 1)

If the only purpose of standards is to "sound the alarm" that too many students are falling below desired levels of achievement, then simple quantifications or cut scores alone would be sufficient. However, if standards are intended to give meaning to the assessment and show what students should be able to do, then more substantive performance standards are needed. To set expectations, standards should embody or represent what students must know to attain the standard. On NAEP, the achievement level descriptions and accompanying exemplar items were intended to convey what was expected at each level.

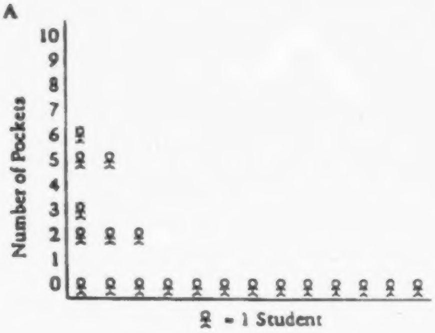
The mathematics problem in Figure 2 is offered as a positive example of an assessment item that exemplifies the intended performance level, in this case fourth-grade advanced. The panel of mathematics experts convened as part of the NAE evaluation identified the "pockets" problem as a good exemplar item. It is consistent with the NCTM content standards because it demonstrates

Overall Percent
Satisfactory or Better
Grade 4:10 (0.9)

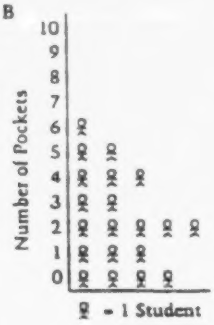
Conditional-Advanced
Grade 4:59%

There are 20 students in Mr. Pan's class. On Tuesday most of the students in the class said they had pockets in the clothes they were wearing.

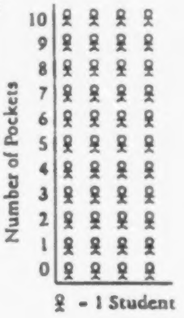
A



B



C



Which of the graphs most likely shows the number of pockets that each child had? B

Explain why you chose that graph.

Because it shows 20 students and most of the students have pockets.

Explain why you did not choose the other graphs.

It cannot be A because in A most of the students do not have pockets.

It cannot be C because in C there are more than 20 students shown.

Figure 2. A good exemplar item used to represent advanced achievement for NAEP fourth-grade mathematics.

Note. Reprinted from *Setting Performance Standards for Student Achievement: A Report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: An Evaluation of the 1992 Achievement Levels* (p. 116), by L. Shepard, R. Glaser, R. Linn, and G. Bohrnstedt, 1993, Stanford, CA: National Academy of Education. Original source: Mullis, I. V. S., Dossey, J. A., Owen, E. H., and Phillips, G.W. *NAEP 1992 Mathematics Report Card for the Nation and States.* Washington, DC: National Center for Education Statistics, 1993.

mathematical reasoning and requires students to communicate mathematically by explaining their answers. Problems like this one go beyond abstract descriptions to show educators and members of the public what performance is desired to meet the standard.

Unfortunately, the 1992 achievement levels in mathematics and reading had several substantive deficiencies that made them inadequate for reporting assessment results. First, in mathematics there were both substantive and technical flaws in the selection of exemplar items. In contrast to the positive example cited above, Figure 3 shows a not-so-ideal sample item also used to exemplify fourth-grade advanced performance. Showing items like this one did not help educators or the public understand the substance of national content standards in mathematics. The use of poor examples, and, ultimately, difficulties with the descriptions themselves, were a direct result of limitations in the NAEP item pools. In a systematic study of the math item pool, content experts found that only 60% of the items matched at least one theme of the NCTM Standards; more seriously, the content validity expert panel concluded that only 9% of the items were truly exemplary in terms of professionally accepted content standards.

In mathematics, there were also statistical problems with some of the exemplar items. If judges picked sample items to represent a given level, but students at the intended level could not, in fact, do the item, then the so-called exemplar item did not really show what students could do. One such item is shown in Figure 4. Because measuring an angle with a protractor is a basic mathematical skill, judges classified the item as basic. However, only 37% of students at the basic level could do the item, so technically it was not a good example of basic student performance. In reading, lessons learned from mathematics were used to correct the problem of statistically misfitting exemplar items. However, there were still problems with the item pool, especially at the advanced level, making it difficult either to assess or represent aspects of the substantive descriptions, such as being able to critically evaluate what has been read.

Ideally, items used to communicate to the public what attainment at each level means should meet both conceptual and statistical criteria. That is, judges should be able to specify desired performance at a level, and, having attained that level, students should (by definition) be able to do what that level requires. Satisfying strict selection criteria, however, does not mean that exemplar items should be such an odd subset that they no longer represent the rest of the items in the pool. Obviously, such a practice would misrepresent the assessment students actually took. As I emphasize again later, the only way to have representative items that can satisfy both substantive and technical criteria is to build the assessment in the first place with the kinds of items and tasks needed to represent knowledge at each of the intended levels.

In 1992, serious problems also arose with the achievement level descriptions. Although treated in much greater detail in the full NAE report, these boiled down to two basic difficulties. First, the descriptions used by the judges to set cut scores were revised substantially after the fact, so, in the end, there was no way of knowing to what extent the cut scores really aligned with the descriptions used to report results to the public. Second, in both reading and math, the motivation for revising the descriptions was to bring them more in line with subject-matter experts' knowledge of national content standards. However, improving the descriptions moved them further away from the existing item pools which had not been developed to reflect current content standards.

If □ represents the number of newspapers that Lee delivers each day, which of the following represents the total number of newspapers that Lee delivers in 5 days?

A 5 + □

B 5 x □

C □ : 5

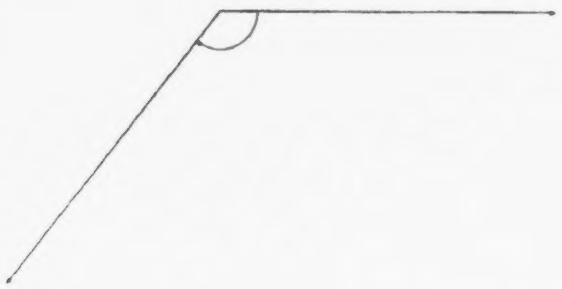
D (□ + □) x 5

Overall Percent Correct*
Grade 4: 48 (1.2)

Conditional-Advanced
Grade 4: 95%

Figure 3. A not-so-ideal exemplar item used to represent advanced achievement for NAEP fourth-grade mathematics.

Note. Reprinted from *Setting Performance Standards for Student Achievement: A Report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: An Evaluation of the 1992 Achievement Levels* (p. 116), by L. Shepard, R. Glaser, R. Linn, and G. Bohrnstedt, 1993, Stanford, CA: National Academy of Education. Original source: Mullis, I. V. S., Dossey, J. A., Owen, E. H., and Phillips, G.W. *NAEP 1992 Mathematics Report Card for the Nation and States.* Washington, DC: National Center for Education Statistics, 1993.



Overall Percent Correct*
Grade 8: 35 (1.9)

Conditional-Basic
Grade 8: 37%

Use your protractor to find the degree measure of the angle shown above

Answer: 128°

*The standard errors of the estimated percentages appear in parentheses.

Figure 4. A so-called "exemplar item" used in reporting NAEP eighth-grade mathematics that does not exemplify what students at the level can do.

Note. Reprinted from *Setting Performance Standards for Student Achievement: A Report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: An Evaluation of the 1992 Achievement Levels* (p. 116) by L. Shepard, R. Glaser, R. Linn, and G. Bohrnstedt, 1993, Stanford, CA: National Academy of Education. Original source: Mullis, I. V. S., Dossey, J. A., Owen, E. H., and Phillips, G.W. *NAEP 1992 Mathematics Report Card for the Nation and States.* Washington, DC: National Center for Education Statistics, 1993.

These mistakes led to the "can" versus "should" dilemma. Ideally, standards are set by stating expectations: what students "should be able to do" to be advanced, proficient, etc.. For example, to earn a lifesaving certificate from the American Red Cross, one must among other things be able to recover a 10-lb weight in 10 ft of water from a 30-ft approach in 9.5 s, be able to swim 50 yds in 33.5 s, and so forth. Upon being certified, the requirements describe what the swimmer "can" do. However, because the 1992 achievement levels were superimposed on an existing assessment and the descriptions altered to reflect national content standards, there were many phrases in the descriptions not tapped by the assessment. (What if candidates for a lifesaving certificate were only asked to swim distances within time limits but were never asked to do the tasks involving weights?) Because important content was left out, reaching the advanced level did not necessarily mean that students could do the things in the advanced description. Therefore, the descriptions were invalid for reporting assessment results.

As suggested earlier, these problems of inadequate substantive standards are relatively easy to fix by starting first with intended, professionally defensible content standards and then developing performance standards. Finally, for each level of performance there must be an adequate array of assessment items or tasks for students to demonstrate all of the aspects of performance expected at that level.

INADEQUACIES OF THE ANGOFF PROCEDURE FOR ELICITING JUDGES' INTENDED STANDARDS

One of the most critical findings of the NAE studies was that the Angoff procedure was inadequate for translating judges' conceptual or substantive standards into cut scores on the NAEP scale. This problem is independent of the earlier problem with descriptive statements not in the assessment, but it also means that the descriptions do not necessarily match up to the score scale as intended. For example, the cut score for fourth-grade advanced mathematics achievement in Figure 1 is 280 (which, by the way, is well above the eighth-grade mean of 268). When the NAE evaluation is quoted as saying that the 1992 standards were "set too high," the criticism does not refer to the substantive or narrative expectations. Rather, the evidence suggests that the cut scores, particularly at the upper end of the scale, were set too high relative to the desired substantive standard and its associated narrative description. As subsequent data demonstrate, the kinds of systematic errors that judges make in judging difficult items and external validity evidence from two separate contrasting-groups studies make it more plausible that the fourth-grade advanced cut score, corresponding to the reported description, should have been around 250 or 260.

The modified Angoff procedure requires that judges participating in the standard-setting process must first envision a hypothetical examinee who just barely meets the conceptual qualifications for a given level. When the Angoff method was first developed, in the context of minimum competency testing, judges had to envision a just-barely-passing, minimally competent knowledge level. For the NAEP application, judges had to imagine a just-barely-qualified basic student, a just-barely-qualified proficient examinee, and a just-barely-qualified advanced examinee. Next, for each level, judges must estimate for every item in the assessment the probability (p-value) that these borderline examinees will get the item correct. For the NAEP achievement levels, judges actually went through three rounds of estimating p-values. After the first round, they were shown data on the actual item difficulties, and after the second round, they were shown which items they were rating the most inconsistently compared to their own intended standard. Each judge's intended standard or cut score was then calculated by summing the third-round probabilities for all the items. Finally, judges' individual cut scores were averaged to arrive at a recommended standard.

Rather than go through all of the internal process studies as laid out in the NAE report Chapter 3 (Shepard et al., 1993), I focus here on the results that prompted the NAE Panel to reconsider the adequacy of the Angoff method and that led ultimately to the conclusion that it is fundamentally flawed. Given the political heat surrounding this conclusion, it is worth noting that before the evaluation began, members of the Panel concurred that the Angoff procedure is a well-accepted, standard-setting approach. For example, in 1980 I personally favored the Angoff procedure compared to other test-centered methods "primarily because it is simpler" (Shepard, p. 453).

The same data that were collected from judges as part of the standard-setting process were reanalyzed to examine the internal consistency of judges' ratings. Table 1 shows one such analysis for mathematics. In each case, the scale score represents the judges' intended cut score for a given level based on different subsets of items. For example, based on easy items, judges recommended a cut score for fourth-grade basic of 187.4; however, for the same intended cut score, when looking at difficult items, judges recommended a cut score of 205.8. Such differences imply that judges were unable to maintain a consistent view of the performance they expected from basic students. For the lay reader, the meaning of these analyses is initially difficult to grasp because it stands to reason that judges would set different expectations or probabilities for easy and hard assessment items. For example, a minimally basic student would have an 80% chance of getting an easy item right but only a 30% chance of getting a difficult item right. However, the scale scores in Table 1 reflect the judges' intended cut scores after the items had been "scaled" to take their relative difficulty into account.

The discrepancies in Table 1 between the intended standard based on easy items and the intended standard based on hard items mean that judges were making judgments that were internally inconsistent and contradictory. The judges apparently could not hold a hypothetical, borderline basic or proficient examinee in mind and estimate a consistent set of p-values for such an individual. It is important to emphasize that the differences in Table 1 are very large. Given a within-grade standard deviation of approximately 40 points, the contradictions at each level range from one-half to one whole standard deviation. These errors are huge if the intention is to carve up the score scale into three or four meaningful categories. Correspondingly, the discrepancies have a big effect on the number of examinees said to have reached each level.

Figure 5 was used in the NAE report to help illustrate how it is that judges' ratings could correlate well with actual item difficulties and still be affected by large systematic errors. Based on the NAEP data and much previous research, judges are reasonably good at rank ordering items by difficulty. However, what has been unexamined in past research is the tendency of judges to compress the scale to a large degree. Thus, the judges systematically underestimate the easiness of easy items but overestimate the easiness of hard items. Other reanalyses in the NAE report also showed that judges were unable to understand differences in the difficulty of multiple-choice versus short-answer items. This again led to internal inconsistencies with judges unable to focus on an intended standard; the apparent inability of judges to take guessing into account especially affected estimates for the basic cut score.

Table I

Mathematics Achievement Level Cut Points, Based Separately on Easy Items (difficulty < median) and Hard Items (difficulty > median), Excluding Extended-Response Items

	Grade 4		Grade 8		Grade 12	
	Scale score	% at or above	Scale score	% at or above	Scale score	% at or above
Basic						
Easy items	187.4	82%	240.2	76%	273.4	76%
Difficult items	205.8	66	270.8	47	302.8	46
Proficient						
Easy items	223.2	46	271.0	47	306.8	42
Difficult items	250.2	16	306.6	14	338.3	12
Advanced						
Easy items	258.2	10	301.6	18	341.4	10
Difficult items	281.0	2	339.2	2	365.5	2

Note. Reprinted from *Setting Performance Standards for Student Achievement: A Report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: An Evaluation of the 1992 Achievement Levels* (p. 59), by L. Shepard, R. Glaser, R. Linn, and G. Bohrnstedt, 1993, Stanford, CA: National Academy of Education.

In rejecting the importance of the NAE findings, a number of commentators have reminded us that it was "just one study." However, it is worth noting here that the systematic inconsistencies in the judges' ratings like those in Table 1 were replicated across three achievement levels, at three different grade levels, in both reading and mathematics. There were no exceptions. Discounting the three achievement levels, which were done by the same panel of judges within a grade, this might more properly be counted as six independent replications of the same finding.

Michael Kane (1995) and other critics have also argued that the apparent inconsistencies might as likely be attributed to problems with the NAEP scaling. The Panel considered problems of multidimensionality and other artifacts of scaling but found them to be implausible as ways to explain the large inconsistencies. First, the scaling method does not involve an indiscriminate combining of all items in the assessment; rather, homogeneous subdomains are scaled and then aggregated to form a composite. Second, the Educational Testing Service (ETS) studies done to examine the effect of item type on dimensionality show distinct but highly correlated factors. Highly correlated item factors could not account for discrepancies of the magnitude found in the NAE studies. For example, multiple-choice and short-answer item factors correlated on the order of .86 to .96. Third, standard errors are too small for the huge hard-easy discrepancies to be attributable to regression. Fourth, the hard-easy comparisons used only items within each grade and, therefore, strained the scaling methodology much less than would occur for vertical equating between grades; yet vertical equating studies have not produced errors on the order of one-half to one standard deviation.

Still, there is a more telling argument against scaling as a counter explanation. The systematic pattern in the judges' misestimation of p-values occurs even when the complexities of Item Response Theory (IRT) scaling are never brought into play. The U.S. General Accounting Office (GAO, 1993) evaluation of the 1990 NAEP math achievement levels found essentially the same pattern of errors at the two extremes using only a comparison of judges' estimated p-values and actual p-values from reported assessment results. The GAO comparisons are shown in Table 2. When setting cut scores for basic students, judges estimated p-values for easy items of only .67, compared to student performance of .83; however, for these same students, judges underestimated the difficulty of hard items, .26 compared to .17. This pattern of judges systematically underestimating the easiness of easy items and at the same time underestimating the difficulty of hard items occurred at every level except for easy items at the advanced level.

The Panel first commissioned these internal reanalyses because of concerns about issues of dimensionality and how current relations among items, based on current instructional practices, might interact with judges' hopes and expectations for the future. For example, the Panel might have expected that judges' intended standards and p-value estimates would correspond closely to current student performance on number and operation items, because this kind of content predominates in current mathematics instruction. In contrast, judges' estimates of what "should be" the standard might depart considerably from current item difficulties in an area like fourth-grade geometry, because such content is not presently being taught. Contrary to these hypotheses, there was no substantive variation in the judges' intended cut scores. At the Joint Conference, Michael Kane (1995) complained that the Panel blamed the Angoff method for everything: first, for the finding of large differences in intended standards and then for the finding of no differences in intended standards across content strata of the assessments. Indeed, the Panel believes that, taken together, these sets of studies reveal what the judges are able and not able to do when asked to make hundreds of p-value estimates. They are able

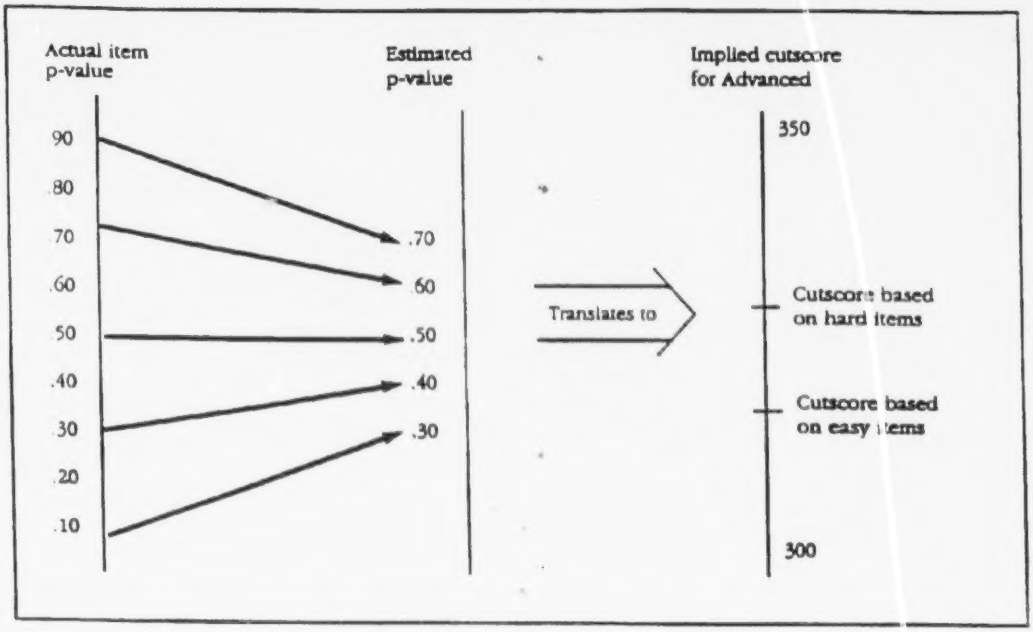

Actual item p-value
.90
.80
.70
.60
.50
.40
.30
.20
.10
Estimated p-value
.70
.60
.50
.40
.30
Translates to
Implied cutscore for Advanced
350
Cutscore based on hard items
Cutscore based on easy items
300

Figure 5. An illustration of a relationship between actual item p-values and judges' estimated p-values with judges selecting less extreme p-values.

Note. Reprinted from *Setting Performance Standards for Student Achievement: A Report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: An Evaluation of the 1992 Achievement Levels* (p. 64), by L. Shepard, R. Glaser, R. Linn, and G. Bohrnstedt, 1993, Stanford, CA: National Academy of Education.

Table II

Comparison of Judges' Item Expectations and 1990 Actual Performance in Mathematics for Items in Each Quartile of Difficulty from the U.S. Government Accounting Office Report (1993)

	Percent correct	
Achievement level and item group	Item judgment expectations	1990 test results
Basic level		
Easy items	67%	83%
Moderately easy items	50	60
Moderately difficult items	40	37
Most difficult items	26	17
Proficient level		
Easy items	85	93
Moderately easy items	75	84
Moderately difficult items	67	67
Most difficult items	56	40
Advanced level		
Easy items	96	96
Moderately easy items	92	95
Moderately difficult items	88	90
Most difficult items	80	71

Note. Adapted from *Educational Achievement Standards: NAGB's Approach Yields Misleading Interpretations.* Report No. GAO/PEMD-93-12 (p. 31 and Appendix VI), 1993, Washington, DC: General Accounting Office. Data source: *The Levels of Mathematics Achievement* (vol. 3, *Technical Report*, pp. 265-71 and vol. 2, *State Results for Released Items*, pp. 3-35), 1991, Washington, DC: National Assessment Governing Board.

to crudely rank-order items by difficulty. However, there is systematic error in their item judgments. Moreover, the lack of differentiation by content area suggests that judges are unable to make nuanced substantive judgments about what "should be" taught and mastered as compared to "what is" presently being taught.

In a separate set of comparisons, the NAE evaluation also found enormous differences between standards set using traditional dichotomously scored items and extended-response items. These differences were the greatest of all, ranging absurdly from four to eight grade levels on the score scale for the same intended cut score. The Panel report did not blame the Angoff procedure, which was used only for the dichotomous items, for this particular problem. The Panel did, however, raise the larger issue that reporting by achievement levels would be meaningless over time as the mix of these items in the assessment changed.

The Panel's conclusion that the Angoff procedure is fundamentally flawed was based on the weighing of evidence from several sources, including the internal consistency data already explained. The evidence that judges tended to underestimate the difficulty of hard items (and, therefore, set cut scores at the upper end of the scale that were too extreme for their intended standard) was consistent with findings from several contrasting-groups studies which showed that more students could actually do the things in the proficient and advanced descriptions than could meet the cut score. This does not mean that the Panel automatically prefers the contrasting-groups methodology; however, when two independent lines of inquiry (internal analyses and external validity comparisons) lead to the same conclusion, their convergence takes on added weight. In addition, a small-scale study where judges set standards using whole assessment booklets instead of item-by-item judgments produced results that were exactly predictable from misestimation of p-values. Using the whole-booklet method, the same group of judges set higher standards for the basic level and a substantially lower cut point for advanced.

Again, a major theme running through the Joint Conference was that the NAE evaluation was just one study and as such should not be used to overturn 20 years of research and practical experience supporting use of the Angoff procedure. But what exactly has been shown by past research? First, the most pervasive finding is that different standard-setting methods produce different results. Absent a certifiable external criterion, such studies have demonstrated that choice of method matters, but they have done nothing to evaluate the validity of the Angoff or any other procedure.

A second major category of research has investigated the reliability of methods. From this kind of study it can be concluded that two randomly equivalent panels of judges, led through the same procedure, will produce acceptably similar results when the judges' estimated standards are averaged. Such reliability studies tell us almost nothing about the substantive integrity of the resulting standards; averaging noise or random error also produces nearly the same average every time. In the case of the NAEP achievement levels, judges disagreed widely with each other, yet their average cut scores were reliable in the sense of these past studies. In 1982, van der Linden conducted a rare study in which he documented internal inconsistencies or "specification errors" in judges' estimates using both the Nedelsky and Angoff methods. With the exception of this study, the measurement literature has not examined whether judges could actually do the task being asked of them consistently.

On the practical side, judges seem to have no difficulty following directions and implementing the Angoff procedure. However, participants' expressions of confidence are very likely influenced by the presence of statistical experts. They trust that their substantive intentions will be validly translated into an appropriate cut score. Judges are never asked, for example, to take an assessment home and see if the derived cut scores indeed correspond to the test scores of real students who most fit their hypothetical just-barely proficient, or just-barely advanced, ideals. Twenty years of successful experiences with the Angoff procedure can also be misleading to the extent that most professional certification and minimum competency programs usually take other information into account, including passing rates, the past year standard, and so forth, rather than basing passing scores automatically on the average of the judges' ratings.

If the nature of evidence is critically examined, it is hard to defend the claim that validity of the Angoff procedure had been established by past research. The NAE evaluation undertook investigations that had never been done before and, based on new evidence, raised serious questions about whether Angoff cut scores really represent judges' conceptual standards.

IMPLICATIONS FOR ALTERNATIVE STANDARD-SETTING APPROACHES

The full NAE report included both short-term recommendations about how assessment results could be reported more meaningfully and long-term recommendations for developing national performance standards congruent with national content standards. For purposes of the Joint Conference, I have focused more on specific standard-setting procedures and what implications can be drawn if judges seem to be unable to think in terms of item p-values.

In addition to the Angoff studies, the Panel convened panels of experts in reading and mathematics and asked them to apply the achievement level descriptions using items mapped directly to the NAEP scale. Judges' experiences when trying to think about standards in this way also provided some valuable lessons about difficulties involved in setting aggregate standards using any item-by-item approach. First, inadequacies in the item pool either made it impossible to locate standards on the scale or forced standards unreasonably high as judges reached higher and higher looking for items that embodied the descriptive standard. Second, the item-mapping approach seemed in retrospect to encourage a false assumption that items and tasks are perfectly correlated--so that an examinee who can do one advanced item is presumed to be able to do all advanced items. This presumption leads to unrealistically high standards and probably accounted for the unreasonable cut scores derived from the boundary exemplar papers as well. Experiences with item-by-item approaches were in marked contrast to the whole-booklet method, where judges were able to consider directly what combination of some advanced items and some proficient items should together constitute advanced performance. Based on logical analysis, as well as these findings, the Panel concluded "that item-by-item methods are inadequate for allowing judges to develop integrated conceptions of performance standards." (Shepard, et al., 1993, p. 78)

The Angoff method or other item-judgment methods might have made sense for setting cut scores for minimum competency exams if content domains were perfectly homogeneous. (Then saying .80 for every item would capture the judges' intention of an 80% passing score for the total test.) However, as assessment domains become more and more complex, it makes less sense that performance standards should be set using a compensatory model of expertise. A compensatory model assumes that any combination of right and wrong answers is as good as any other combination. Instead of a vertical scale, I've drawn a picture in Figure 6 to suggest that, in complex domains, knowledge progressions are better represented by correlated, but diverging vectors or by a branching tree. For example, all measurement experts should have core knowledge regarding correlation coefficients and reliability, but as knowledge becomes more specialized, two experts could know strikingly different amounts about Item Response Theory calibration or measurement policy. The higher the standards and the more heterogeneous the performance domain, the less likely that statistical averaging or straight compensatory models will work. Representing expertise is likely to require a "mixed model," where certain skills and competencies are essential, but where trade-offs are allowed at higher levels of specialization.

Holistic approaches, where judges look at real student performances in combination, guard against false assumptions about the intercorrelation of items and have a better chance of allowing judges to specify what trade-offs they are willing to accept and still judge an overall performance to be advanced. Putnam, Pence, and Jaeger's (1994) efforts to set standards for the National Board for Professional Teaching Standards provided an example where expert judges were involved directly in articulating their intended policies underlying cut score decisions. After being asked to look at an Avogadro's number of hypothetical profiles and participating in a statistical-policy-capturing study intended to infer their policies, judges were asked in a follow-up survey, "Well, what would your policy be?" The experts started saying things like this: "This should be an advanced standard because this is the National Board. But we don't necessarily want all 4s. Is my standard going to be a combination of 3s and 4s? Would I ever allow a 2?" Some were really insightful saying something like, "Well, I'd allow a 2 but not if they ever got two 2s on the same dimension," which obviously reflects an understanding of measurement error. My point is that new approaches to standard setting must allow for this kind of model of expertise and must allow judges to engage in this kind of thinking and reasoning to arrive at a correspondence between their conceptual standard and an appropriate cut score.

Experts on standard setting acknowledge that true standards do not exist in nature like parameters waiting to be estimated by a reliable statistical procedure. Regardless of how statistical they seem, standard-setting procedures are merely formal processes intended to help judges approach their conceptual task in as systematic a fashion as possible. Unfortunately, with hundreds of items to be judged for multiple levels and multiple rounds, judges may be distracted from thinking about the substance of the standards they are trying to set. This imbalance of form over substance needs to be reversed in future approaches so that judges can focus on substance.

Standard setters should be encouraged to identify marker profiles of performance at each level (possibly using simplified, but real composites of student performance). Then they should be asked if they are satisfied with the result in terms of samples of students and items classified at each level. They should not rely on a statistical translation of their work. Efforts being made presently to build the content of assessments to reflect content and performance standards will also help with the problem of setting cut scores; but judgments will still have to be made about how many proficient items and how many advanced items signify attainment at each level.

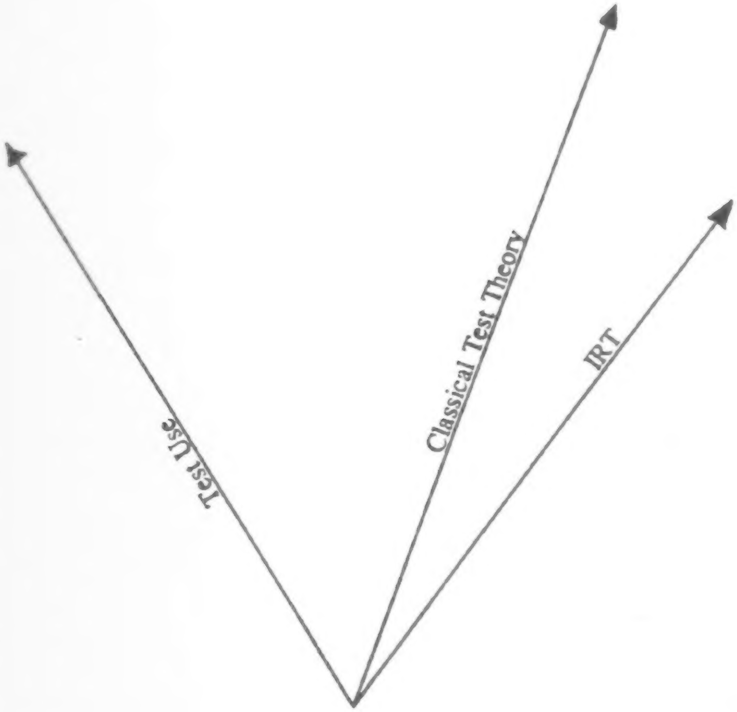


Figure 6. A proposed model of measurement expertise suggesting that knowledge progressions in complex domains are better represented by correlated, but diverging vectors.

References

Bourque, M. L., & Garrison H. (1991). *The levels of mathematics achievement: vols. 1-3. National and state summaries*. Washington, DC: National Assessment Governing Board.

Finn, Jr., C. E., Jr. (1989) [*News* release]. Washington, DC: National Assessment Government Board.

Kane, M. (1995). Examinee-centered vs. task-centered standard setting. In *Proceedings of Joint Conference on Standard Setting for Large-Scale Assessment*. Washington, DC: National Assessment Governing Board and National Center for Education Statistics

Mullis, I. V. S., Dossey, J. A., Owen, E. H., & Phillips, G. W. (1993). *NAEP 1992 mathematics report card for the nation and the states*. Washington, DC: National Center for Education Statistics.

Putnam, S. E., Pence, P., & Jaeger, R. M. (1994, April). A multi-stage dominant profile method for setting standards on complex performance assessments. Paper presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, New Orleans.

Shepard, L. A. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4,* 447-467.

Shepard, L. A., Glaser, R., Linn, R. L., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement*: A report of the National Academy of Education Panel on the evaluation of the NAEP trial state assessment: An evaluation of the 1992 achievement levels. Stanford, CA: Stanford University, National Academy of Education.

U.S. General Accounting Office. (1993). Educational achievement standards: NAGB's approach yields misleading interpretations. (Rep. No. GAO/PEMD-93-12). Washington, DC: Author.

van der Linden, W. J. (1982). A latent trait method for determining intrajudge inconsistencies in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement, 19,* 295-308.

Standard Setting--The Next Generation

Ronald A. Berk[1]

Professor, Johns Hopkins University School of Nursing

ABSTRACT

Specific directions for future standard-setting practices are proffered by answering three key questions: (1) Where have we been? (2) What have we learned? and (3) So what's a practitioner to do? The answers are presented in the form of three "Top 10 Lists." Technical issues related to reliability and validity evidence for the standard-setting process and a standard-setting awards finale complete the presentation.

WARNING:

NAGB, NCES, NAE, GSA, GAO, OMB, and FDA have determined that this material may contain harmful or dangerous ideas. This is the only time in history when this group of initials has agreed on anything!

CAUTION:

DO NOT drive or operate dangerous machinery while reading this material.
If **DROWSINESS** occurs, it shouldn't be surprising.
DO NOT induce vomiting; the content will take care of that.

Truckloads of articles, pamphlets, and book chapters have been published on the topic of standard setting. When charged with the task of executing a particular standard-setting method, executors throughout the land have exhibited high blood pressure, high anxiety, ulcers, and, in one case, even hairballs. After two decades of research and practice, the measurement community is a smidgen closer now than it was then to grappling with the pesky and prickly complexities of standard setting.

At present, there are nearly 50 standard-setting methods documented in the literature. At the epicenter of every method proposed since prehistoric times is "human judgment," whose subjectivity and imprecision wreak havoc in the minds of quantitatively-trained people--which might explain some of the grumbling and murmuring among standard-setting researchers at professional conferences. The various methods require judgments about test content (the items or the test itself), examinees' performance, or a combination of both. You may be thinking: "How soon can we expect to see an objective, unbiased, nonjudgmental method?" I think we have a better chance of seeing nutritional information on the side of a Slurpee or seeing the Cincinnati Bengals win the Super Bowl! If you're still clinging to this dream of a nonjudgmental method, I suggest: Let It Go! In fact, what were previously labeled as judgmental and empirical methods were renamed by Jaeger (1989) as *test-centered* and *examinee-centered* methods, respectively, to acknowledge that all methods are judgmental.

---

[1] The author gratefully acknowledges the assistance of Ellen Spies in the preparation of this paper and of all the materials for the presentation upon which the paper was based.

Recently, the test-centered methods used by the National Assessment of Educational Progress (NAEP) to set achievement levels for the 1990 and 1992 assessments were the objects of considerable controversy and, in fact, a heavy barrage of criticism (Cizek, 1993; Kane, 1993; Linn, Koretz, Baker, & Burstein, 1991; Shepard, Glaser, Linn, & Bohrnstedt, 1993; National Academy of Education [NAE], 1993; Stufflebeam, Jaeger, & Scriven, 1991; U.S. General Accounting Office, 1993). The simplest single-stage methods, such as Angoff (1971), Ebel (1979), and Nedelsky (1954), that were extremely popular and reportedly effective tools for setting cut scores in a variety of applications in the 1980s and early '90s, are now being pinned to the mat. The NAE (Shepard et al., 1993) recommended that NAEP's modification of the Angoff method, as well as any other item-judgment method be discontinued in favor of the contrasting-groups approach (Livingston & Zieky, 1982), which should be "used to evaluate the current set of achievement levels" (p. xxx). The reason given was that such methods are fundamentally flawed because they require judges to perform a nearly impossible cognitive task, i.e., estimate the probability that hypothetical students at the borderline of each achievement level will answer each item in the pool correctly.

When I first read this recommendation, it blew me off my beach chair. The Angoff method with all of its modifications has been one of the most trusted names in standard setting. And now it seems as though it's fighting for its life along with the other item-judgment methods. However, despite the specific advantages and disadvantages of each cut-score method, which have been debated extensively, NAE (Shepard et al., 1993) concluded that:

> The most consistent finding from the research literature on standard setting is that different methods lead to different results. Not only do judgmental and empirical methods lead to different results, . . . but different judgmental methods lead to different results. In fact, judgmental methods appear to be sensitive to slight and seemingly trivial differences in the procedures used to implement a given method. (p. 24)

Jaeger's (1989) analysis of 32 comparisons of methods led to virtually the same conclusion. Probably the only point of agreement among standard-setting gurus is that there is hardly any agreement between results of any two standard-setting methods, even when applied to the same test under seemingly identical conditions. This conclusion is based on empirical evidence no less.

So, where does the research go from here? There are at least four alternatives to consider. The first alternative is to pick a method that seems to fit the requirements of a specific application and then bite the bullet and simply acknowledge its psychometric limitations. A second choice is to select one method, but analyze the range of cut scores by the judges, rather than just the average, in conjunction with external validity evidence to guide the choice of a cut score. Unfortunately, it is unclear how these different types of information should be integrated to arrive at a final standard. Third, several fearless, standard-setting leaders (Hambleton, 1980; Jaeger, 1989; Shepard, 1984) have suggested using two or more methods in a given application and then considering all of the results, together with extraterrestrial factors, when determining a final cut score. Since it is reasonable to forecast that these results will not converge on the same cut score, how does one reconcile the difference? This strategy might create somewhat of an unsettling feeling among a few practitioners. Finally, there is one other not so obvious option. The components of all methods and the available evidence of their effectiveness can be scrutinized in order to build an eclectic approach that capitalizes on the best and the brightest elements in standard-setting history. Of course, this approach would have to be tested, but it might be worth the effort.

To provide a different direction for future standard-setting practices in large-scale assessments, I have chosen the deviant course of constructing an eclectic model. However, in order to embark on this construction job, two key questions need to be answered: (a) Where have we been? and (b) What have we learned? Once these have been answered, one can answer the most critical question: So what's a practitioner to do?

Since there is some evidence that my previous papers on standard setting have produced drowsiness and, in rare instances, nausea and vomiting, readers should take heed of the caution on the cover of this paper. In an effort to minimize the possibility of these side effects, the vehicle for presenting this material will be three "TOP 10 LISTS." This strategy also capitalizes on the "THREE" phenomena in history. Several historians, all of whom wish to be disassociated from this paper, have observed that many great events and turning points in history have occurred in "THREES," such as the Three Musketeers, measures of dispersion (range, variance, standard deviation), and, of course, Angoff-Ebel-Nedelsky. This paper will emulate that distinguished model.

WHERE HAVE WE BEEN?

Numerous summaries of the standard-setting literature have been reported throughout the 1980s (Berk, 1986; Hambleton, 1980; Hambleton & Eignor, 1980; Jaeger, 1989; Pulakos, Wise, Arabian, Heon, & Delaplane, 1989; Shepard, 1980, 1984), and one review appeared in the '90s (Kane, 1994). Most of these method-by-method critiques are quite lengthy. In fact, I can be accused of writing one of the longest pieces on the topic, and I have regretted the pain I inflicted on the readers of my "consumer's guide" for quite some time. In order to redeem myself, I recently proposed the shortest summary of standard setting ever recorded (Berk, 1995). A revision of that summary will be presented here.

There is a remarkable similarity in structure among the approximately 30 methods that utilize a test-centered, judgmental review process, which has been the most popular, feasible, and credible approach in many large-scale assessments to date. While there are numerous variations in what judges rate, how they rate, and the steps required to reach the cut score, the research and experience with these methods have taught more about what works in the standard-setting process than have the less frequently used and studied examinee-centered, empirical methods. Therefore, the summary presented here is derived primarily from the test-centered methods and, to a lesser extent, from the examinee-centered methods. Rather than reviewing the specific methods, which has already been done, this presentation lists the pieces in the standard-setting puzzle that have been extracted from those methods.

Top 10 List Number One

So, from the home office in Baltimore, Maryland, here are "Berk's Top 10 Characteristics of the Judgmental Standard-Setting Process" (a.k.a. Where Have We Been?):

10. **Given enough time, judges can be trained to do just about anything.**

9. **Judges can rate objective or constructed-response items, clusters of items, work samples, profiles of behavior, or whole tests.**

8. **Judges can weight the importance, difficulty, or complexity of objectives or behavioral dimensions.**

7. **Judges can match the content in any item format or an examinee's item/work sample/test performance to behavioral descriptions of achievement levels.**

6. **Judges can use feedback on individual ratings, panel ratings, and/or pertinent quantitative performance information on examinees after their initial ratings.**

5. **Judgments about standards can be determined by judges independently or through consensus after group discussion.**

4. **The judgmental process can be single-stage, multistage, or iterative.**

3. **The pattern of ratings structured by objective or dimension can be compensatory, conjunctive, or disjunctive.**

2. **The standard can be computed based on consensus of the panel of judges or from the average or median percentage/proportion across judges, weighted average proportion across judges, median percentage across samples of judges, or any of those measures in conjunction with an examinee performance-based scale.**

And the number 1 characteristic of the judgmental standard-setting process:

1. **The standard that emerges from any method will probably be adjusted higher or lower based on judgments about the political, economic, social, and/or educational consequences of the decisions.**

These characteristics can be viewed as ingredients or options available in the process. They have been applied in a variety of combinations since the early 1970s to test scores used in criterion-referenced or competency tests of student performance; teacher and professional licensing and certification exams; personnel evaluation of teachers, counselors, and administrators; and performance appraisal of employees and managers in private industry and the military. The pivotal question is: What optimal combination of ingredients produces an effective standard-setting procedure? As I have already suggested: We don't know! The problem is that the measurement community has never reached consensus on a set of criteria that can operationally define the "effectiveness" of any single standard-setting method. That's what makes research on this topic so much fun! There are NO RULES!

Lists of criteria for evaluating the various methods have been proposed previously (see, for example, Berk, 1986; Hambleton & Powell, 1983; Plake, 1995). The 1985 edition of the *Standards for Educational and Psychological Testing* (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) provides standards that focus more on the technical outcomes in terms of reliability and validity evidence, such as Standards 1.23, 1.24, 2.10, 2.12, and 11.3, than on the standard-setting process itself. Only Standard 6.9 states that the method, rationale, and technical analyses should be reported along with the qualifications of the judges involved in the process:

> When a specific cut score is used to select, classify, or certify test takers, the method and rationale for setting that cut score, including any technical analyses, should be presented in a manual or report. When cut scores are based primarily on professional judgment, the qualifications of the judges also should be documented. (*Primary*) (p.49)

Although these published sets of criteria and the *Standards* do not provide the level of specificity the profession seems to need to guide standard-setting practices, there are high expectations among many measurement professionals that the current revision of the 1985 *Standards* will set explicit requirements for the entire standard-setting process.

WHAT HAVE WE LEARNED?

With no professionally agreed-upon criteria at this time, how can the "best" elements or attributes of the standard-setting process be determined? The previous Top 10 List presents a confusing array of options. Furthermore, techniques that were effective during the 1980s may not be applicable now. Throughout the '80s, item-judgment methods, such as Angoff's and its numerous variations, were the predominant choices of practitioners for setting standards in a variety of applications. At the outset of the 1990s, two trends in the large-scale assessment business seriously questioned the utility of these "most-loved" methods: (a) the increased use of multipoint item formats and (b) the requirement of multiple cut scores. Either the older methods had to be retooled or new approaches had to be developed. Following is a discussion of some of the recent developments in these areas.

Recent Developments

Multipoint Item Formats

The performance assessment mania sweeping the country has resulted in a meteoric rise in the use of constructed-response item formats, essays, writing samples, oral discourses, exhibitions, experiments, portfolios, and assessment centers for student, teacher, and administrator examinations (Berk, 1993; U.S. Congress, Office of Technology Assessment, 1992). This increased usage has translated into a variety of scoring protocols and multipoint or polytomous (a.k.a. polychotomous) items. Exercises that are scored on a range of 0-2 or greater have necessitated a different cut score strategy from the item-probability judgments recommended by most standard-setting methods for dichotomous items (see, for example, Hambleton & Plake, 1995).

New approaches are currently being investigated to move beyond the simple item-by-item judgments of yesteryear. These approaches include the following: (a) setting cut scores on each multipoint item and the total test (Delaware Department of Public Instruction, 1993; Freed, 1993), (b) setting scale score bands or intervals around achievement level thresholds (Atash, Ferrara, & Bayless, 1994; Kahl, Crockett, & DePascale, 1994), (c) requesting judges to specify score performance distributions (Poggio & Glasnapp, 1994), and (d) stating decision policies across an entire assessment package (Putnam, Pence, & Jaeger, 1995). These efforts concentrate judgments on scores derived from aggregates of items, exercises, or work samples. Yet they also capitalize on many of the aforementioned elements that have proven successful in past standard-setting projects, such as using an iterative process (Jaeger, 1982), and are equally applicable to the traditional dichotomously-scored item format. However, further research is needed to evaluate the most effective "elements" in these applications.

Multiple Cut Scores

In another twist on the passing score theme, there has been a trend by a few agencies to set multiple cut scores, such as two, three, or four, to categorize examinees by levels of achievement, competency, or proficiency. For example, NAEP specifies "Advanced," "Proficient," and "Basic" (American College Testing, 1993; National Assessment Governing Board, 1990); the National Board for Professional

Teaching Standards recommends "Highly Accomplished," "Accomplished," "Competent," and "Novice" (Jaeger, 1995); the Kansas State Board of Education uses "Excellent," "Strong," "Progressing," "Borderline," and "Inadequate" (Poggio & Glasnapp, 1994). Several other state education agencies, including Delaware, Kentucky, Maine, Maryland, Massachusetts, and New Hampshire, have proposed similar schemes (see, for example, Atash et al., 1994; Delaware Department of Public Instruction, 1993; Freed, 1993; Kahl et al., 1994; WESTAT, 1993).

As if setting one passing score weren't challenging enough to defend psychometrically, these multiple cutoffs complicate the judgmental process even more. This complication begins with defining every level of achievement for which a cut score must be determined. The achievement level definitions serve as behavioral descriptions or examples of performance. The procedure used to describe these levels is referred to as "behavioral anchoring." Selected points on a scale, called anchor points, are described in terms of the knowledge, skills, and abilities (KSAs) exhibited by students who are near those points. These descriptions are similar to the "benchmarks" used on behaviorally-anchored rating scales (BARS) (Smith & Kendall, 1963) and behavior-summary scales (BSS) (Borman, Hough, & Dunnette, 1976) originally designed to rate the job performance of employees in industrial, educational, and governmental organizations in the early 1960s and '70s (for examples of scales, see Borman, 1986). An example of behavioral anchoring for four achievement levels is shown in Table 1. Once the achievement levels are explicitly defined, anchor or prototype exercises, behaviors, work samples, or other item formats are assigned to the designated levels by the judges.

This strategy for defining three or more achievement levels and sorting student behaviors into these levels has been tested in conjunction with different Item Response Theory (IRT) methods in several state assessment programs. For example, the one-parameter partial-credit model has been used to scale test scores, item-difficulty values, and examinees' ability estimates onto a common scale called an "activity difficulty scale." Researchers at the Maryland State Department of Education (Atash et al., 1994; WESTAT, 1993) have explored the use of a judgmental process to set achievement levels on this difficulty scale with narrow bands of points that discriminate between performance at two adjacent levels. The behaviors and skills assigned to these levels by the judges provide the foundation for deriving descriptions of the achievement levels.

Other IRT approaches have been investigated by Kahl et al. (1994) in Maine's and New Hampshire's assessment programs. A Student-Based Constructed Response (SBCR) method, where students are placed on a Rasch ability scale based on their scores on all of the common questions they answered, requires judges to review a complete set of responses for every student whose work they examine. Alternatively, an Item-Based Constructed Response (IBCR) method places score points for individual items on the ability scale. In this method, judges review responses sorted by score point by item. Both the SBCR and IBCR methods request judges to participate in "range-finding" activities to minimize their review load. They match student work to predetermined definitions of four levels of achievement ("Distinguished," "Proficient," "Apprentice," "Novice") and three cut scores on the scaled score continuum.

Table I

Example of behavioral anchoring for 4 achievement levels

	ULTRASTUPENDOUS (Gold Medal)	
Leaps tall buildings in a single bound.	Is more powerful than a locomotive.	Is faster than a speeding bullet.
	KINDA BETTER THAN ORDINARY (Silver Medal)	
Leaps short buildings in two or more bounds, maybe.	Can draw a picture of a locomotive.	Is about as fast as a speeding BB.
	MIDDLING GARDEN-VARIETY (Bronze Medal)	
Leaps short buildings with a running start and a strong tailwind.	Can pick out the locomotive in an H-O train set.	Tries to catch speeding BBs in teeth.
	PUTRID (Pet Rock)	
Barely leaps over a Port-O-Potty.	Says "Look at the choo-choo!"	Wets self while shooting a water pistol.

Other variations of the behavioral-anchoring method have been explored by the National Board for Professional Teaching Standards. Jaeger (1995) and Putnam et al. (1995) investigated two "judgmental policy-capturing" strategies adapted from a procedure developed by Hobson and Gibson (1983) for the performance appraisal of employees. This procedure focuses on the pattern of responses or the policy used by judges to arrive at their final standard. The judgmental process involves defining multiple levels of competency, rating profiles of teacher performance, and then rating hypothetical candidates with specified profiles of exercise scores from "Novice" to "Highly Accomplished." The policy that emerges to describe the judges' responses can take one of three possible forms: (a) compensatory, where the cut score is simply the total score across all items and objectives; (b) conjunctive, where objectives/dimensions are weighed by importance and different cut scores may be set across objectives in order to pass; or (c) disjunctive, where a mixture of compensatory and conjunctive may be designated with cut scores set for some objectives and total score computed across other objectives.

This analysis of how judges' decisions are made by item and objective to produce a single cut score or pattern of cut scores for use with a test was considered for the student competency and teacher licensing tests in the 1970s and '80s. However, the compensatory model with its inherent simplicity (i.e., just add 'em up) reigned supreme. The research by the National Board has refocused attention on these decision rule options to more accurately and sensitively reflect how judges think and feel about what candidates should be able to perform.

Top 10 List Number Two

Considering all of the developments within the last four years in conjunction with the previous decade of research and experience with item-judgment methods, here are "Berk's Top 10 Picks for the Judgmental Standard-Setting Process" (a.k.a. What Have We Learned?). *(Note:* The following picks are for professional discussion only and should not be used as the basis for any actual cash wager.)

10. **Select a broad-based sample of the most qualified and credible judges you can find.**

9. **Train these judges to perform the standard-setting tasks to minimize "instrumentation effect" and maximize "intrajudge consistency."**

8. **Use a multistage iterative process whereby judges are given one or two opportunities to review/refine their original decisions based on new information to maximize "interjudge consistency."**

7. **Require judges to provide explicit behavioral descriptions for each achievement level with corresponding anchor items.**

6. **Determine the judges' decision policy based on the objectives or dimensions measured.**

5. **Provide judges with feedback on their individual and the panel's decisions.**

4. **Supply judges with meaningful performance data on a representative sample or appropriate subsample of examinees to reality-base the ratings.**

3. **Allow judges the opportunity to discuss their decisions and pertinent data without pressure to reach consensus.**

2. **Solicit judges' content-related decisions about achievement levels via consensus, but all item and test score decisions via independent ratings to avoid pressuring "outlier judges" into alignment or the influence of "dominant judges."**

And the number 1 pick for the standard-setting process:

1. **Compute the cut score(s) from the mean/median item or test scores based on the judges' ratings.**

SO WHAT'S A PRACTITIONER TO DO?

If the elements in this preceding Top 10 List could be synthesized into a package plan to be manufactured as a generic brand of standard setting, what would it look like? It should incorporate the best of the past with the most promising new techniques in order to handle any item or test format and single or multiple cut scores, and yet be feasible for large-scale assessments. The structure should permit the user to choose between test-centered and examinee-centered components, and it should be applicable to student as well as teacher assessment. The steps in this Generic Eclectic Method (GEM) are listed next.

Preliminary Steps

Two giant steps need to be completed before the standard-setting games begin: (a) select samples of judges and (b) train the judges.

Select Samples of Judges

Pick the most credible judges you can dig up. For most large-scale assessments it is advisable to choose two samples of judges: (a) a broad-based, diverse group representative of the population of educators, such as teachers, curriculum experts, and local and state-level administrators; and noneducators, such as parents, professionals in a variety of disciplines, and the general public; and (b) a specialized group of content experts by grade level corresponding to the test content areas and grade levels. Beyond the professional composition of the samples, both should be representative of all sociodemographic characteristics. This assures sampling precision as well as political correctness and avoids the appearance of an all-dweeb judgmental panel. The broad-based sample should be selected first; then the content experts can be drawn as a subsample or as a separate sample.

The broad-based rather heterogeneous sample should be employed to make decisions about achievement-level definitions and items within the limits of its expertise. For tests measuring specialized content across a range of cognitive levels at different grade levels, one or more homogeneous subsamples should be comprised of appropriate content experts with the qualifications to render valid decisions at the various steps in the process. The composition of each subsample of judges should accurately reflect the types of content and levels of cognitive complexity being measured, the difficulty of the items, and the grade levels. Unqualified judges who are required to make uniformed decisions at any stage of the process will tend to contribute inaccurate information and error variance. The validity and generalizability of the decisions made by the judges depend on the judges' competence, representativeness, and credibility.

BEST COPY AVAILABLE

Jaeger's (1991) suggestions regarding sample representativeness and size should be considered in this selection process:

> Judges should be selected through procedures that permit generalization of their collective recommendations to well-defined populations. The number of judges . . . should be sufficient to provide precise estimation of the standard that would be recommended by an entire population of judges, . . . such that the standard error of the mean or median recommended standard is small, compared to the standard error of measurement of the test for which a standard is sought. (p. 10)

Train the Judges

Train these judges 'til it hurts. The effectiveness of the training program can determine the effectiveness of the entire standard-setting process. The major objectives of the training are to provide judges with a thorough understanding of the concepts of achievement or proficiency levels and competence in the tasks required by the standard-setting method. The research on factors that influence judges' ratings during training and the stages of standard setting deal almost exclusively with item judgments and defining minimal competence (Melican & Mills, 1986, 1987; Mills, Melican, & Ahluwalia, 1991; Norcini, Shea, & Kanya, 1988; Plake, Melican, & Mills, 1991; Pulakos et al., 1989; Smith & Smith, 1988). Despite the limited focus of these studies, the criteria proposed by Reid (1991) for evaluating training effectiveness based on their findings can be generalized to the more recent standard-setting methods: (a) judgments should be stable over time, (b) judgments should be consistent with item and test score performance, and (c) judgments should reflect realistic expectations.

When conducting the training, how much practice is enough? For example, suppose judges are matching nontest items from the item domain to previously defined achievement levels. Each judge should be able to establish a high level of consistency (intrajudge) in matching items at different levels of difficulty or complexity to the appropriate achievement levels. The judges should not waiver during the process in their understanding of the items, the achievement levels, or in the matching itself. Once judges have reached this level of competence in the specific matching or rating task and are confident in the stability of their decisions, SOCK IT TO 'EM!

Since the background characteristics of the judges can affect the speed with which each judge masters the task, individualized training using an interactive computerized presentation should be considered. A computerized procedure will also permit constant monitoring of the consistency of the judges' performance during training and facilitate intrajudge consistency during the standard-setting process (Plake et al., 1991). The judges' confidence in their decisions can also be assessed and correlated with their consistency. A rating scale can be used to measure confidence at the conclusion of training and after the final decisions have been made during standard setting. (To reward the judges for all of their efforts during training, take them to lunch or buy gifts for their children!)

Top 10 List Number Three

Are you ready for this? Another TOP 10 LIST. Talk about beating a dead horse! For those of you who watch "Letterman," you'll probably never watch him again! Well, here comes "Berk's Top 10 Steps to Standard-Setting Salvation" (a.k.a. Generic Eclectic Method or GEM), plus criteria to evaluate each step:

10. **Broad-based sample of judges defines achievement levels by subject area and grade level based on *consensus*.**

CRITERIA: This diverse group of judges should reach *consensus* on general policy and content definitions of what types of examinee performance are expected at each designated achievement level. If policy definitions have already been set by a governing board, then either the judges can extend those definitions to clarify their intent or skip this step entirely, pass GO, and collect $200.

9. **Subsample (content experts) of judges provides amplified, explicit behavioral descriptions of achievement levels by subject area and grade level based on *consensus*.**

CRITERIA: This merry band of content experts should build on the general definitions produced in Step 10. They should reach *consensus* on specific, operational descriptions of the knowledge, skills, and abilities (KSAs) that must be demonstrated at the different achievement levels. The meaning of those levels and the interpretation of the final cut scores hinge on the clarity of the behavioral definitions at this step. (The descriptions will most likely need revisions once the entire process has been completed. The actual items selected in subsequent steps of the process may suggest some refinements in the original behavioral descriptions to improve their meaning.)

8. **Judges select anchor/prototypic unscored items (test-centered) or scored items/work samples/tests (examinee-centered) at upper and lower ends of achievement level categories based on *consensus*.**

CRITERIA: This step is the standard-setting version of "bootcamp," which will give judges the opportunity to BE ALL THAT THEY CAN BE! It is intended to train the judges in item content and difficulty, achievement level definitions, and the matching process. Nontest items should be used for this practice exercise until competence and confidence are attained. This training may be conducted on an individualized basis via computer.

At this step, decisions must be made on the unit of judgment (item, work sample, or total test), item scoring format (dichotomous, polytomous, or a combination of both), and test-centered (unscored unit) or examinee-centered (previously scored unit) approach. For example, the units may consist of multiple-choice and constructed-response items that have not been administered to any examinees. Matching the content and perceived difficulty of these items to the achievement levels is a test-centered approach. Alternatively, if the units have already been administered and scored based on a sample of examinees, then the scored level of performance on each item becomes part of the matching criteria. In other words, since the examinees' scores are used to assign the items to the achievement levels, the method is examinee-centered.

Once all judges have completed the training, they should be convened to select anchor items/work samples from the practice pool for upper and lower ends of the achievement levels. These anchors should provide concrete representations of the descriptions and include all item formats used on the final test. The judges should reach consensus on a set of anchors for each achievement level. (The anchors may need relocation or replacements after they have been tested in the process.)

7. **Judges *independently* match unscored items (test-centered) or scored items/work samples/tests (examinee-centered) to achievement level categories based on behavioral descriptions and anchors.**

CRITERIA: After the training and final agreement on anchors at Step 8, each judge *independently* should match all test items/work samples/tests to the appropriate achievement levels based on content, perceived difficulty, and, if scored, level of performance. If this process could be conducted on microcomputers, each judge could be monitored to detect "drift" in the degree of interjudge inconsistency (Plake et al., 1991). For unscored multipoint items, judges should select item cut scores (e.g., three out of four) corresponding to the achievement level classifications.

6. **Judges *independently* rate the importance, difficulty, or complexity of each objective/dimension.**

CRITERIA: If the test is structured by objectives or other content categories/ dimensions, an absolute rather than relative scale should be used to weight the objectives to allow any possible decision policy to emerge. A range of anchors for the importance scale should be presented for each objective, for example, "Essential," "Extremely Important," "Very Important," "Somewhat Important," "Not Very Important," and "Unimportant."

5. **Judges are given feedback on their individual and the panel's decisions plus meaningful performance data *AND* requested to *independently* revise their initial decisions.**

CRITERIA: Feedback should occur in the forms of item distributions by achievement levels and estimated cut scores based on those distributions for each judge and across the panel. Performance data from a representative sample or appropriate subsample of examinees should include item difficulties, IRT scale item values if applicable, and estimated score distributions and cut scores based on the panel's decisions. In addition, the percentage of examinees at and above each cut score should be presented to communicate the potential impact or consequences of the cut score decisions. Such information may help judges understand their classifications and ratings in terms of the results. If gaps exist between what they intended and the projected cut scores, the judges then have the opportunity to eliminate those gaps by adjusting their ratings. (*Note:* Depending on whether a test-centered or examinee-centered approach has been used, p-values or an IRT difficulty scale may be introduced to the judges at Step 7 or 8 when the items are matched to the achievement levels. There is a sparsity of evidence on whether such data are useful to judges earlier in the process [Atash et al., 1994].)

Once all of this information has been presented to the judges--on a microcomputer, if possible--they should be encouraged to revise the initial item or work sample classifications and importance ratings. Individual decisions will maintain confidentiality and privacy in the judgmental process.

4. **Judges *discuss* their item or work sample classifications and importance ratings without pressure to reach consensus.**

CRITERIA: This open discussion will provide the only opportunity for the judges to interact and process each other's explanations and justifications for their decisions. This step may help judges clarify discrepancies in their own classifications and ratings and affirm or disaffirm their decisions. (It should be noted that for this discussion session, dress is usually casual, although a few judges have shown up wearing lederhosen.)

3. **Judges render their final *independent* revisions of their classifications and ratings based on the discussion (Step 4), as well as accumulated insights.**

   CRITERIA: As a follow-up to the discussion and all prior information, each judge should be able to privately make any final adjustments in classifications and ratings without the direct influence of the panel. This step will terminate the judges' involvement in the process. (Before they disband, throw them a going-away party; invite their families to your house for the holidays!)

2. **Determine the judges' decision policy rule from the designated weights finally assigned to the objectives/dimensions (Step 3).**

   CRITERIA: If the judges' median weights of importance across objectives do not yield a distinctive conjunctive or disjunctive pattern, then the compensatory model should serve as the default. Certainly all objectives rated as equally important would indicate a compensatory policy as well.

And the number 1 step to standard-setting salvation:

1. **Compute the cut scores for the total test or by objective/dimension based on the decision policy rule (Step 2) and the mean/median item or test scores from the judges' final classifications at each achievement level (Step 3) in terms of the chosen test score scale.**

   CRITERIA:
   1. **Compensatory policy**--Single or multiple cut scores for the total test are computed from the judges' mean/median scores.
   2. **Conjunctive policy**--Cut scores are computed from the judges' mean/median scores by objective.
   3. **Disjunctive policy**--Cut scores are computed from the judges' mean/median scores for most important objectives and aggregated for less important objectives.

This 10-step process is an iterative, behavioral-anchoring approach. The unit of judgment may be the item in any format, a cluster of items, a work sample such as a portfolio, or the total test. Judges should not be requested to dream up item probabilities for any hypothetical sample of examinees. They should be asked to draw on their content expertise to provide behavioral descriptions of one or more achievement levels, select anchor items/work samples/tests for those levels, and then to match the test items or the test itself to the levels. If the unit being matched has been scored previously (examinee-centered), those scores represent cut scores corresponding to the achievement levels (in the minds of the judges). The mean/median score for the sample of judges at the end of the process will be the cut score for a particular level.

If the unit of judgment has not been scored (test-centered), either the point value of the item or separate cut scores on the multipoint items can be summed and averaged across judges to establish cut score levels. Alternatively, IRT scaled scores based on the assignment of items or work samples to difficulty values (examinee-centered) can be used to transform the judges' decisions into cut scores at the specific achievement level(s).

Decisions must be made on six issues in order to apply this generic method to a specific cut score situation: (a) examinee target population (students, teachers, administrators, etc.), (b) unit of judgment (item, cluster of items, work sample, or total test), (c) item scoring format (dichotomous, polytomous, or a combination of both), (d) test-centered (unscored unit) or examinee-centered (previously scored unit or IRT scale) approach, (e) number of achievement levels or cut scores, and (f) weighting of objectives for decision policy analysis. The two iterations at Steps 3 and 5 are designed to refine the judges' decisions and improve the likelihood of a high degree of interjudge consistency at Step 3.

WHAT'S WRONG WITH THIS STANDARD-SETTING PICTURE?

So far, not a single quantitative index of reliability or estimate of misclassification error has been mentioned. Time's up! Reliability and validity evidence related to the judgmental process and the consequences of the decisions should be collected to determine whether the method works.

Reliability Evidence

There are three indices that can be computed to assess the degree to which the judges behaved themselves or, at least, behaved according to the psychometric expectations: (a) intrajudge reliability between steps, (b) intrajudge reliability within steps, and (c) interjudge reliability. Although these indices will be discussed separately here, the analysis of the different sources of measurement error can be conducted in one generalizability study, where the variance components for each source can be isolated and indices calculated for a given application.

Intrajudge Reliability Between Steps

The iterative judgmental process recommended in the Top 10 Steps may yield high intrajudge variance across the steps (or occasions) inasmuch as the procedure encourages each judge to revise, refine, and reconsider decisions based on new information (Step 5) and input from other judges (Step 4). This process affords the judges two chances to change their earlier item/work sample classifications and ratings of objectives. Variability across the three sets of decisions (Steps 7, 5, and 3) can be expected to be relatively high if judges are integrating the information presented and making the appropriate adjustments in their ratings. Consequently, an estimate of intrajudge reliability across the steps of the judgmental process should yield a low coefficient.

Intrajudge Reliability Within Steps

The estimate of intrajudge reliability that is of greater concern is the consistency within Steps 8, 7, 5, and 3. Key elements in the process are intended to maximize this type of consistency: (a) requesting judges to provide explicit behavioral descriptions of the achievement levels to improve clarity in interpretation (Step 9), (b) training judges to match items/work samples/tests to achievement levels at acceptable levels of consistency and confidence (Step 8), (c) using anchor items to improve the accuracy of item-achievement level classifications (Step 7), (d) providing judges feedback on their own decisions and performance data to adjust decisions to be consistent with realistic expectations (Step 5), (e) permitting discussion of decisions among judges for clarification (Step 4), and (f) iterations at

Steps 3 and 5 for judges to revise or refine their decisions. Other strategies to improve intrajudge consistency have been identified by Plake et al. (1991).

Reliability checks should be conducted during and after training (Step 8), after the first formal item matching process (Step 7), and after the final decisions are rendered (Step 3). One cannot assume that high levels of consistency attained in the early stages will be sustained throughout the process. Estimates need to be computed at several checkpoints to monitor each judge's consistency.

Interjudge Reliability

In contrast to the above sources of intrajudge variance, the estimate of interjudge reliability measures the degree of homogeneity or internal consistency of the final decisions by the judges (Step 3). It is essentially an index of convergence. However, despite the iterations and specific steps in the process designed to increase the likelihood of convergence, interjudge consistency may be relatively low due to any number of factors (e.g., ambiguity in definitions of achievement levels, format of items or exercises, competence of judges, background characteristics of judges). Mean/median scores can still be translated into cut scores for the objectives and total test in the presence of considerable interjudge variability.

A generalizability coefficient computed from the between-judges variance component would furnish evidence of the dependability of the cut score(s) generated from the specific sample of judges. It would indicate the extent to which the standard is replicable or generalizable to other samples. This is an important criterion for evaluating the success of the process.

These three measures of "reliability" reflect the characteristics of the process itself; that is, they are based on the internal mechanism for generating the standard. They should not be confused with the other indices of reliability reported in the literature (Berk, 1984) and required by Standards 2.10, 2.12, and 11.3 (AERA, APA, NCME Joint Committee, 1985), including decision consistency and standard errors of measurement, which are external to the judgmental process. Those indices are calculated from the score distributions of examinees on one or more occasions.

Validity Evidence

The internal validity of the process hinges on the qualifications of the judges and the procedure used to solicit their judgments. After digesting Berk's Top 10 Lists and scrutinizing every procedure that can be used in the standard-setting process, it can be concluded that the final standard is "whatever the judges say it is." This is certainly not a compelling argument for validity evidence, but the credibility of the group of content experts and procedural fidelity are the only available internal criteria. What is missing in this rather circular thinking is an external criterion (Berk, personal communication while jogging, August 1994; I talked to myself).

Consequential

No matter how well the internal mechanism is polished, only external evidence can provide insights into the consequences of the decisions or whether the correct decisions were made. Step 1 of the first Top 10 List states how consequences have been incorporated into standard setting. However, judges can be given the opportunity, as illustrated in Step 3 of the process (Top 10 List Number Three), to consider the consequences of their decisions based on estimated cut scores and the percentages of examinees who obtained them. Adjustments at this step make it possible to contextualize the judges' decisions in terms of consequences before the final standard is determined.

Any standard emerging from the judgmental process can be adjusted to account for a variety of what Messick (1989) refers to as "social consequences and side effects of the decisions" (p. 21). Evaluating these consequences, intended or unintended, of student and teacher testing is essential to the validation of the decision-making uses of test scores. Criterion measures of the political, economic, social, and/or educational outcomes of decisions about examinees must be obtained to determine whether or to what extent cut scores should be raised or lowered. This analysis of consequences should encompass all pertinent issues, such as school improvement and adverse impact.

Evidential

In addition to evidence related to the consequential basis of test use, the evidential basis of test interpretation and score use must be addressed. Specifically, evidence of the success of candidates who take the test can indicate the accuracy of the decisions made based on the cut score(s). Decision validity or accuracy (Hambleton, 1980) is the acid test of the worth of a standard-setting method. Many of the empirically based standard-setting methods of the 1970s and '80s required the collection of this type of evidence in the form of traditional hit rates (a.k.a. probability of correct decisions) and false-positive and false-negative errors. Such evidence is also stipulated by Standard 1.24 (AERA, APA, NCME Joint Committee, 1985).

Decision validity evidence is based on predictive studies of the relationship between performance status on the test (e.g., proficient-basic) and actual performance in a subsequent position (Kane, 1982). For professional licensing and certification examinations, defining the criterion of success and measuring it without bias or distortion have been the bane of most test-criterion relationships. Messick (1989) has identified numerous sources of bias that can "operate differentially on the three major forms that criterion measures typically take: namely job samples, supervisors' ratings, and production records in industrial settings or achievement tests, teacher judgments, and career data in the case of educational outcomes" (p. 73).

Despite these psychometric limitations and the real-world roadblocks to gathering predictive evidence, a commitment to investigate the accuracy and meaningfulness of the decisions made on the basis of cut scores is long overdue. Achievement levels at the upper grade levels, in particular, can be linked to entry-level job performance data and success in college. Linkage models developed in military applications (Wise, 1994) can be adapted to large-scale assessment standards. Studies to determine the relationships between achievement levels in different content areas and the skills from a representative sample of jobs should be conducted to evaluate some of the consequences of pass-fail decisions.

AN AFTERTHOUGHT OR TEN

What implications do the preceding Top 10 Lists and challenges to collecting reliability and validity evidence have for future standard-setting practices? I don't have a clue! (Just kidding!) Actually, probably more is known about what doesn't work than what does. Although you probably don't want to read anything that's packaged in "10s" by now, my summary of the main points of this paper is an awards presentation, expressed in (what else?) a "Top 10 **MOST** Awards List":

10. **Most Promising Old Approach:** Multistage Iterative Process

9. **Most Promising I/O Psych Approach:** Behaviorally Anchored Scaling

8. **Most Politically Correct Procedure:** Selecting a Broad-Based Panel of Judges

7. **Most Politically Incorrect Procedure:** Using Content Experts Only to Set Standards

6. **Most Confusing New Term:** Polychotomous or Polytomous

5. **Most Psychometrically Incorrect Procedure:** Asking Unqualified Judges to Make Uninformed Decisions

4. **Most Challenging Old Procedure:** Consensus Building Among Judges

3. **Most Challenging New Complication:** Setting Multiple Cut scores

2. **Most Neglected Technical Topic:** Evidential (Predictive) Validity

1. **Most Difficult to Defend:** All of the Above!

As Sergeant Phil Esterhaus of "Hill Street Blues" used to say to his officers every morning: "Hey, Let's Be Careful Out There!"

References

American College Testing. (1993). *Setting achievement levels on the 1992 National Assessment of Educational Progress in mathematics, reading, and writing*: A technical report on reliability and validity. Iowa City: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Atash, M. N., Ferrara, S., & Bayless, D. (1994, June). *A new method for setting absolute standards*. Paper presented at the National Conference on Large-Scale Assessment sponsored by the Council of Chief State School Officers, Albuquerque, NM.

Berk, R. A. (1984). Selecting the index of reliability. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 231-266). Baltimore: Johns Hopkins University Press.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56,* 137-172.

Berk, R. A. (1993). National trends in student and teacher assessment: Issues in performance assessment. In National Evaluation Systems, *Performance assessment in teacher certification testing* (pp. 17-33). Amherst, MA: Author.

Berk, R. A. (1995). Something old, something new, something borrowed, a lot to do! *Applied Measurement in Education, 8*(1), 99-109.

Borman, W. C. (1986). Behavior-based rating scales. In R. A. Berk (Ed.), *Performance assessment: Methods and applications* (pp. 100-120). Baltimore: Johns Hopkins University Press.

Borman, W. C., Hough, L. M., & Dunnette, M. D. (1976). *Development of behaviorally based rating scales for evaluating the performance of U.S. Navy recruiters* (Tech. Rep. TR-76-31). San Diego, CA: U.S. Navy Personnel Research and Development Center.

Cizek, G. J. (1993). *Reactions to National Academy of Education report, Setting performance standards for student achievement*. Washington, DC: National Assessment Governing Board.

Delaware Department of Public Instruction. (1993). *Delaware interim assessment reading and mathematics technical report*. Dover: Author.

Ebel, R. L. (1979). *Essentials of educational measurement* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Freed, C. W. (1993, April). *The standard-setting process for the interim assessment period: A white paper*. Dover: Delaware Department of Public Instruction.

Hambleton, R. K. (1980). Test score validity and standard-setting methods. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art* (pp. 80-123). Baltimore: Johns Hopkins University Press.

Hambleton, R. K., & Eignor, D. R. (1980). Competency test development, validation, and standard setting. In R. M. Jaeger & C. K. Tittle (Eds.), *Minimum competency achievement testing: Motives, models, measures, and consequences* (pp. 367-396). Berkeley, CA: McCutchan.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*(1), 41-55.

Hambleton, R. K., & Powell, S. (1983). A framework for viewing the process of standard setting. *Evaluation and the Health Professions, 6,* 3-24.

Hobson, C. J., & Gibson, F. W. (1983). Policy capturing as an approach to understanding and improving performance appraisal: A review of literature. *Academy of Management Review, 8,* 640-649.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4,* 461-476.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). New York: American Council on Education/Macmillan.

Jaeger, R. M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice, 10*(2), 3-6, 10, 14.

Jaeger, R. M. (1995). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education, 8*(1), 15-40.

Kahl, S. R., Crockett, T. J., & DePascale, C. A. (1994, June). *Using actual student work to determine cut scores for proficiency levels: New methods for new tests*. Paper presented at the National Conference on Large-Scale Assessment sponsored by the Council of Chief State School Officers, Albuquerque, NM.

Kane, M. (1982). The validity of licensure examinations. *American Psychologist, 37,* 911-918.

Kane, M. (1993). *Comments on the NAE evaluation of the NAGB achievement levels*. Washington, DC: National Assessment Governing Board.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64,* 425-461.

Linn, R. L., Koretz, D. M., Baker, E. L., & Burstein, L. (1991). *The validity and credibility of the achievement levels for the 1990 National Assessment of Educational Progress in mathematics* (CSE Rep. 330). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Melican, G. J., & Mills, C. N. (1986, April). *The effect of knowledge of item difficulty for selected items on subsequent ratings of other items using the Angoff method*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Melican, G. J., & Mills, C. N. (1987, April). *The effect of knowledge of other judges' ratings of item difficulty in an iterative process using the Angoff and Nedelsky methods*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Mills, C. N., Melican, G. J., & Ahluwalia, N. T. (1991). Defining minimal competence. *Educational Measurement: Issues and Practice, 10*(2), 7-10.

National Academy of Education. (Shepard et al., 1993). *Setting performance standards for student achievement*. Stanford, CA: Author.

National Academy of Education. (1993). *Setting performance standards for student achievement: Background studies*. Stanford, CA: Stanford University, National Academy of Education.

National Assessment Governing Board. (1990). *Setting appropriate achievement levels for the National Assessment of Educational Progress: Policy framework and technical procedures*. Washington, DC: Author.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Norcini, J. J., Shea, J. A., & Kanya, D. T. (1988). The effect of various factors on standard setting. *Journal of Educational Measurement, 25*(1), 57-65.

Plake, B. S. (1995). The performance domain and the structure of the decision space. *Applied Measurement in Education, 8*(1).

Plake, B. S., Melican, G. J., & Mills, C. N. (1991). Factors influencing intrajudge consistency during standard setting. *Educational Measurement: Issues and Practice, 10*(2), 15-16, 22, 25.

Poggio, J. P., & Glasnapp, D. R. (1994, April). *A method for setting multi-level performance standards on objective or constructed response tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Pulakos, E., Wise, L. L., Arabian, J., Heon, S., & Delaplane, S. K. (1989). *A review of procedures for setting job performance standards*. Washington, DC: American Institutes for Research.

Putnam, S. E., Pence, P., & Jaeger, R. M. (1995). A multistage dominant profile method for setting standards on complex performance assessments. *Applied Measurement in Education, 8*(1).

Reid, J. B. (1991). Training judges to generate standard-setting data. *Educational Measurement: Issues and Practice, 10*(2), 11-14.

Shepard, L. A. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4,* 447-467.

Shepard, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 169-198). Baltimore: Johns Hopkins University Press.

Smith, P., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47,* 149-155.

Smith, R. L., & Smith, J. K. (1988). Differential use of item information by judges using the Angoff and Nedelsky procedures. *Journal of Educational Measurement, 25,* 259-274.

Stufflebeam, D. L., Jaeger, R. M., & Scriven, M. (1991). *Summative evaluation of the National Assessment Governing Board's inaugural 1990-91 effort to set achievement levels on the National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board, August 1991.

U.S. Congress, Office of Technology Assessment. (1992, February). *Testing in American schools: Asking the right questions* (1992-297-934Q13). Washington, DC: U.S. Government Printing Office.

U.S. General Accounting Office. (1993). *Educational achievement standards: NAGB's approach yields misleading interpretations* (Rep. No. GAO/PEMD-93-12). Washington, DC: Author.

WESTAT. (1993, June). *Establishing proficiency levels and descriptions for the 1992 Maryland School Performance Assessment Program*. Rockville, MD: Author.

Wise, L. L. (1994). Setting performance goals for the DOD linkage model. In B. F. Green & A. S. Mavor (Eds.), *Modeling cost and performance for military enlistment* (pp. 37-60). Washington, DC: National Academy Press.

Standard Setting: The Next Generation

Summary of Break-out Session[1]

The session began with a discussion of *reliability and the training of judges*. There was consensus in the group that: (a) the training of judges is critical, (b) training must be focused on *process* and not on the complexity of content, and (c) monitoring of training through iterative judgmental steps is vital.

In response to a question about the iterative process of rating items in the training of judges, Berk first discussed "instrumentation effect," warning against changing the concept of the training process for judges. He explained that in a step-by-step process of iteration, judges can work on nonactual items, and, from a high level of variability between steps, they can reach a level of stability that supports the process of intrajudge consistency. He noted that once judges start the real process, there is no change between steps; the expectancy is that, by then, intrajudge inconsistency should be low. He emphasized that judges can be trained to match items in a reliable, unchanging way, but cautioned that the training program can be sabotaged if definitions are not clear from the beginning; he did not consider item format to be an issue in training judges. One participant disagreed with the latter statement, claiming that inferences cannot be made from multiple choice items, for instance. Berk maintained his position; however, he underscored that whatever the item format, it is extremely important that judges reach a level of consistency.

The group then discussed *how to convey results to the public*. Opinions were varied. One participant suggested that descriptions of significant percentage categories of students could be published and that samples of items and answers could be displayed at schools' open house programs and on other similar occasions. Another suggested that reports with supplementary information could be provided. This information could include a description of tasks, definitions of skill levels, and percentage of students at each level. It was suggested that other more detailed information, along with students' work samples, might also be shared with the public.

The group discussed the high level of public interest in *high-stakes testing programs*. This interest, it was suggested, stems in part from issues related to standards and levels of performance. An example is the discrepancy between high standards and low performance. It was suggested that, before reporting results to the public, some issues--such as the labeling of performance levels and the number of performance levels--need to be resolved and must be clearly articulated. One participant observed that value-laden terms should be avoided in the reports.

The group then turned its attention to a discussion of the importance of *change over time*. It was noted that, at the state level, it is important to show gradual improvement in test scores over time. It was suggested that maintaining the accuracy of norm descriptions is important. Berk pointed out the difficulties of reporting results at one point in time. He indicated that reporting change over time would have its difficulties as well, and he suggested that to do this, it is necessary to consider two areas: (a) labeling the level of improvement or change to effectively inform the public, and (b) the period of time over which to maintain the same standards.

---

[1] This is a summary of the break-out discussions for Ronald Berk's presentation. The session was facilitated by Jeanne Griffith (National Center for Education Statistics) and recorded by Patricia Dabbs (National Center for Education Statistics).

An Interpretive Approach to Setting and Applying Standards

Pamela A. Moss[1]

Assistant Professor, University of Michigan, School of Education
Educational Studies Program

ABSTRACT

Interpretive research traditions suggest alternatives to assessment practices developed within the psychometric tradition. They also provide perspectives from which mainstream assessment practices can be critiqued. In this paper, I compare psychometric and interpretive approaches to setting and applying standards in the context of high-stakes individual assessment. Perhaps the most striking contrast between the two approaches involves the ways in which standards are typically applied to individual cases. Within the psychometric approach, single performances are evaluated independently, the scores algorithmically combined, and the composite score compared to a pre-established cut score. Within the interpretive approach, readers consider all the available evidence on an individual, evaluating each part in light of the whole, and engage in a consensus-seeking discussion to reach a decision about whether standards have been met. In comparing these approaches, I consider the quality of information provided, the potential consequences to various stakeholders, and the implications for developing and maintaining an assessment system under each approach.

Both interpretive researchers and psychometricians are concerned with constructing and evaluating interpretations based on human products or performances. Major differences between interpretive and psychometric approaches to assessment can be characterized largely in terms of how each treats the relationship between the parts of an assessment (individual products or performances) and the whole (entire collection of available evidence) and how human judgment is used in arriving at a well-warranted conclusion. Within the psychometric approach, single performances are evaluated independently, the scores algorithmically combined, and the composite score compared to a pre-established cut score. These scores, whose interpretability and validity rests largely on previous research, are provided to users with guidelines for interpretation. In contrast, within the interpretive approach, readers consider all the performances available on an individual, evaluating each part in light of the whole, and then engage in a consensus-seeking discussion to reach a decision about whether standards have been met. The goal is to construct a coherent interpretation of the collected performances, continually revising initial interpretations until they account for all of the available evidence. Here, the interpretation might be warranted by criteria like readers' extensive knowledge of the context; multiple and varied sources of evidence; an ethic of disciplined, collaborative inquiry which encourages challenges and revisions to initial interpretations; and the transparency of the trail of evidence leading to the interpretations, which allows users to evaluate the conclusions for themselves.

Although the focus of this paper is on issues of setting and applying standards, it is necessary to locate these issues in a more comprehensive description of an interpretive assessment process; so the discussion is somewhat broader than the title reflects. To illustrate interpretive approaches to assessment, I will focus on two contexts of assessment where high-stakes decisions about individuals are made: teacher certification and student certification for high school graduation. In the first context, I suggest an alternative relevant to national or state assessment of teachers; in the second context, I draw on existing descriptions of interpretive practices already in place at some high schools. In comparing psychometric and interpretive approaches, I consider questions regarding the quality of information provided, the potential consequences for various stakeholders, and the implications for maintaining an assessment system.

---

[1] I am grateful to Gail Baxter for helpful comments on an earlier draft of this paper. The paper was written while the author was working under a fellowship from the National Academy of Education/Spencer Postdoctoral Fellowship Program.

For each context, the interpretive alternative I outline can be thought about in one of three ways. First, it can be taken at face value as a viable option for serving the purpose of teacher or student certification. As indicated above, in the context of student certification for graduation, a number of high schools have already begun using such an approach. Second, the interpretive alternative can provide a point of comparison or triangulation for construct-validity research within a more typical approach to assessment. If the two approaches lead to similar decisions, then the validity of each is enhanced; if they lead to different decisions, then it becomes important to understand the reasons for the difference, and again knowledge is enhanced. Finally, the interpretive alternative can provide a perspective from which more typical assessment practices can be critiqued (and of course, vice versa). When we consider alternative possibilities--alternative means of serving the same purpose--the practices we take for granted become more visible and open to change. At the very least, we become more aware of the consequences of the choices we make. In this way, the comparison I draw is an exercise in Messick's (1989) Singerian mode of inquiry, where one inquiring system is observed or evaluated in terms of another inquiring system to highlight the assumptions and values underlying each.

EXAMPLE 1: TEACHER CERTIFICATION

For the first example, I outline an interpretive approach to assessment for teacher certification and compare it to a more typical approach based in psychometrics. The example draws on the set of exercises and scoring dimensions developed by the English/Language Arts Assessment Development Lab of the National Board for Professional Teaching Standards (Delandshere & Petrosky, 1992, 1994; Jacobson & Stilley, 1992). The outline of psychometric and interpretive alternatives covers, first, the actual scoring process, where standards are applied and, second, the process of setting standards. (Readers should note that in comparing psychometric and interpretive approaches to setting and applying standards, I am not intending to represent the practices planned by the National Board. I am simply using the exercises and scoring dimensions to build two realistic but hypothetical scenarios.)

As described by project staff, there were 11 exercises on which candidates' credentials were to be based. Three of the exercises, representing the school-site portfolio, were to be completed by the candidate over a period of months at their school sites. They consisted mainly of videotapes or other artifacts of their teaching, accompanied by extensive written reflection. The artifacts about which candidates wrote included a videotape of students discussing a piece of literature; lesson plans and actual activities for a 3-week unit of instruction together with a videotape of teaching; and samples of writing collected from two different students over a period of three months. The content examination, completed on site at an assessment center, contained four written exercises covering candidates' knowledge of literature, of reading and writing processes, and of language development. Finally, four additional exercises, also completed at the assessment center, asked candidates to participate in a small-group discussion to select a reading list for students, to analyze a videotape of an instructional episode, to plan a unit of instruction on language diversity based on materials provided, and to evaluate pieces of student writing provided at the assessment center. To evaluate the exercises, the assessment creators developed six separate dimensions, a subset of which was to be applied to each exercise. The dimensions included learner centeredness, cultural awareness, content knowledge, integrated curriculum, coherent pedagogy, and professional concerns.

That rich set of performances lends itself to scoring, setting standards, and reaching decisions in both interpretive and psychometric ways. It is sufficiently standardized to permit an aggregative approach to scoring and yet sufficiently flexible to encourage candidates to demonstrate their particular strengths. At this point, I present two hypothetical scenarios about how these performances might be used to arrive at a well-warranted credentialing decision: an "aggregative" approach based in psychometrics and an "integrative" approach based in the interpretive research tradition.

Scoring Performances

Within an *aggregative* approach to scoring, human judgment is limited to single performances which are algorithmically combined and compared to a cut score. This is not to imply that human judgment is not involved in developing the method of combining scores or setting the cut score. However, in the actual scoring, the decisions are algorithmic once the exercise scores have been determined. Readers, after they are trained, work independently. They do not discuss the performances in question; rather, their judgments are also combined algorithmically, although sometimes a criterion reader adjudicates disparate scores.

It is useful to question both the quality of information and the consequences of using assessment designed under these two approaches. For instance, Wineburg (1988), drawing on cases from the Teacher Assessment Project (TAP) at Stanford, offers the following examples where information is lost when exercises are scored independently. The TAP researchers were developing prototype assessments for teacher certification, involving a rich and varied set of performances from each candidate. Although the intent was to score them separately, Wineburg approached two cases holistically, reading the entire set of performance available on those candidates. He described a situation where one candidate had referred in detail to the same historical event across three different exercises. The candidate received high scores for content knowledge on each exercise. In another situation, a candidate used the term "critical thinking" in an interview. When probed, the candidate did not elaborate. And so, the candidate received a low score on the relevant dimension. However, on an earlier exercise, an interview about a videotape of the candidate teaching, not only did the candidate elaborate on the concept of critical thinking but executed it brilliantly in interactive teaching. These are both instances in which scoring exercises independently would lead to misinterpretations.

In contrast, the goal of a more *integrative* approach is to construct a coherent interpretation based on the entire set of performances, continually revising initial interpretations until they account for all the available evidence. Here, readers work together, debating their interpretations of the actual performances, and seeking consensus on an interpretation or decision. Nothing about the process is algorithmic; human judgment is involved at every stage. In higher education, useful analogies can be found, for instance, in the dissertation exam or a tenure decision.

Table 1 presents a detailed comparison of the two approaches to scoring performances and to setting standards. In the *aggregative* approach, each exercise is scored by a different pair of readers, working independently, on a predesignated subset of six scoring dimensions (learner centeredness, cultural awareness, content knowledge, integrated curriculum, coherent pedagogy, and professional concerns). The readers are blind to the candidate's performance on all other exercises. Readers provide a score on a criterion-referenced scale specifying level of accomplishment. These scores are (weighted and)

Table 1
Comparison of Aggregative and Integrative Approaches

	An aggregative approach	**An integrative approach**
Scoring procedure	Readers trained to score one exercise	Readers trained to score entire set of exercises
	Different readers evaluate each exercise independently	Single pair of readers evaluates entire set of exercises for a candidate: first alone, and then in dialogue
	Different exercises evaluated on different scoring dimensions	All exercises evaluated on all scoring dimensions (i.e., relevant evidence noted wherever it appears)
	For each exercise, each reader produces a score on designated scoring dimension(s)	For each exercise, readers identify and record evidence relevant to all scoring dimensions
	Scores are (weighted and) aggregated across exercises and/or dimensions following an algorithm	Reader pairs jointly produce evidence matrix, compiling evidence from each exercise and dimension, and prepare clinical commentary
Applying standards	Operationalized through cut scores determined by one or more of the typical standard-setting procedures	Operationalized in reader training by comparing exemplars (performances, evaluations, and explanations) of successful and unsuccessful candidates
	Overall decisions reached by comparing candidate's score(s) to cut score(s)	Overall decision reached through discussion of reader pairs who then jointly prepare written justification, tying evidence to decision
Reader reliability	Evaluated by exercise and dimension, having different readers independently score samples of performances	Evaluated by overall decision, having different pairs of readers independently score samples of candidates' performances (content analysis and comparison of evidence matrices could also be conducted on a sample)
Audit/ confirmation	(Performances receiving disparate scores may be evaluated/adjudicated by a criterion reader)	Overall decision, written justification, and evidence matrix audited by criterion reader: EITHER for random sample of candidates OR for all candidates OR for candidates where reader pairs are not confident of decision
Feedback to candidate	Candidate receives overall decision, profile of scores, and guidance on interpreting scores	Candidate receives overall decision and written justification along with clinical commentary

aggregated into score scales, by scoring dimension or exercise, and possibly into an overall score. The score(s) are then compared to one or more cut scores derived from judgmental standard-setting procedures, and a credentialing decision is made. Interreader reliability is monitored at the exercise level by dimension, through generalizability analyses. Feedback to the candidate might consist of a statement of the decision, a profile of scores (by dimension, exercise, or both), and an explanation of how to interpret those scores.

In the *integrative* approach, a single pair of readers evaluates the entire set of performances from a candidate. The goal is to construct a coherent and well-warranted interpretation of the candidate's teaching knowledge, practice, and reflection, using the scoring dimensions as a framework to structure the interpretation and then to evaluate that interpretation in light of the performance standards set by the credentialing agency. Individual readers first work through the exercises alone, noting and recording evidence on any of the six dimensions wherever it occurs. Then they work together to compile their evidence into a jointly produced matrix that summarizes the evidence available for each dimension and provides clinical commentary on each exercise to be included in feedback to the candidate. (See Figure 1 for a suggested template.)

Concerns about readers reaching premature judgments after encountering responses to only a few exercises might be dealt, in part, with an ethic of inquiry that encourages readers to try alternative interpretations and seek disconfirming evidence. Readers then debate and reach consensus on a credentialing decision, perhaps at a yes or no level, or perhaps on a more refined scale indicating level of accomplishment. Readers then jointly prepare a written justification, tying the evidence they have gathered to the conclusion.

The performance standards cannot be articulated in a cut score. Rather, they are operationalized in reader training by comparing exemplars (performances, evaluations, and explanations) of successful and unsuccessful candidates, as determined by a committee of accomplished teachers. Exemplars include all original performances from a candidate, evaluations, and explanations of those evaluations. This approach implies, of course, a substantially more extensive training period for readers, although actual reading time might be equally efficient. In the case of teacher certification, this may suggest that rather than bringing in many readers for brief and intensive scoring sessions, the credentialing agency would hire accomplished teachers as fellows for 1 or 2 years.

The decisions, written justification, and supporting evidence are audited/confirmed by a criterion reader, perhaps for all the candidates, perhaps for a random sample of candidates, or perhaps for those candidates where the reader pairs do not have high confidence in their decisions. As part of the ongoing monitoring of the system, in addition to confirmatory readings, a sample of candidates' performances is independently rescored by a second pair of readers and the consistency of the decisions evaluated. Comparative content analyses of the evidence matrix and written justifications could help in understanding disagreements and in designing future training sessions for readers.

		Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5	Dimension 6	Clinical Commentary by Exercise
SCHOOL SITE PORTFOLIO	Exercise 1							
	Exercise 2							
	Exercise 3							
CONTENT EXAM	Exercise 4							
	Exercise 5							
	Exercise 6							
	Exercise 7							
ASSESSMENT CENTER	Exercise 8							
	Exercise 9							
	Exercise 10							
	Exercise 11							
Clinical Commentary by Dimension								

Figure 1. Reduced template for evidence matrix.

Candidates receive copies of the decision, the written justification, and the clinical commentary. Thus, the candidates receive expert feedback, tailored to their performances, and are able to trace for themselves the trail of evidence leading to the decision. They have the right to appeal the decision, perhaps asking for an audit by a criterion reader or for an independent second reading by a different pair.

Setting Standards

Within a psychometric or *aggregative* approach, the setting of performance standards, operationalized in cut scores, is often based on information that has been (at least partially) abstracted from the original performances. Members of a standard-setting committee, typically accomplished professionals within the area of assessment, are asked to consider information like scoring rubrics or profiles of scores. Although they may be intimately familiar with sample performances typical of those the rubrics or score profiles represent, the debate and decisions often occur around these more decontextualized pieces of information.

Within an interpretive or *integrative* approach, the setting of performance standards would be based on complete cases; and the standards would be characterized, in large part, through exemplars of cases representing different levels of performance. An exemplar would consist of all performances available on a candidate; a performance level or certification decision reflecting the consensus arrived at by a standard-setting committee; and a justification, citing the evidence that led to the decision. To set standards, individual cases comprising complete sets of performances on each individual would be presented to a committee of highly accomplished professionals within the relevant assessment domain. In creating exemplars of different levels of performance, standard-setting committees would read all the materials available on a candidate, compile evidence, and debate the relevant decision. Written notes and transcriptions of the recorded debate would be used to compile the case-based justifications. After a series of such exemplars had been developed, a more general description of satisfactory performances could be induced from the justifications. That description, along with the exemplars, could be shared with the relevant policy-setting body for approval. The bank of exemplars could be expanded and the description revisited periodically as new exemplars are added and the standards evolve to keep up with current views of best practice.

This practice of setting and applying standards is analogous to the practice of law, where laws are interpreted and applied, through development of precedents (or case-based exemplars), which in turn inform the specific application to new cases. Also, like the law, the standards could evolve over time and the exemplar bank remain active as new cases are added and some cases reinterpreted or deleted (overruled).

Neither of the approaches to assessment outlined above is fully characteristic of the epistemological principles that underlie psychometrics or interpretive methods. In the aggregative approach as described, there is opportunity for discussion and debate in various aspects of the assessment process: as exercises and scoring rubrics are designed, as standards are set, as research is planned, and so on. (The original scoring plans described by Delandshere and Petrosky, 1992, even involved discussions between readers who jointly produced scores and interpretive summaries for individual exercises. Such strategies are more typical of an interpretive approach.) The proposed integrative approach involves readers who are unfamiliar with the context in which the candidate works and who have been trained to adopt a predetermined set of standards. These choices, which limit the interpretive nature of the proposed approach, are necessary in the current assessment context. Given our definitions of fairness, it would be inappropriate for the credentialing decisions to depend upon who the readers are.

A Validity Research Agenda

A program of research that compares interpretive and psychometric approaches within different contexts of assessment would be extremely valuable to the profession. As a construct validity study, these two approaches could be applied independently to the same sample of candidates' performances. (At the least, such a comparison would be an informative means of evaluating the results of standard-setting efforts.) Appropriately designed, the comparison would inform questions of reader reliability and decision consistency within each approach, of time involved in training readers and scoring materials, and of decision consistency across approaches. If the different approaches resulted in consistent certification decisions, then the validity of each would be enhanced; if they resulted in substantially different decisions, then it would become important to understand and explain the differences. If feasible, the sample of candidates could be drawn from groups known, through other direct means, to reflect varying levels of accomplishment. That would allow comparison of three means of identifying accomplished candidates and address the question of which approach to assessment does the better job of distinguishing among the known groups. Finally, the results and feedback from both approaches could be shared with the candidates in the sample to see which they found most useful and credible.

EXAMPLE 2: STUDENT CERTIFICATION FOR GRADUATION

In this second example, I focus on the context of student certification for graduation. Here, there are a number of schools around the country where interpretive approaches, frequently described as "graduation by exhibition," are already being used. Case studies of such schools are provided, for example, by Ancess, Darling-Hammond, and Einbender (1993); Davis and Felknow (1994); Mabry (1992); and McDonald, Smith, Turner, Finney, and Barton (1993). The procedures for applying and setting standards and for warranting the validity of certification decisions outlined above in the context of teacher certification are also applicable to the context of student certification. In this section, I describe a typical process and highlight issues relevant to the consequences of different assessment choices.

In some schools, the process resembles a dissertation exam. Students prepare a portfolio typically consisting of work exhibits that demonstrate their capabilities with respect to school or state curriculum standards, or both, along with reflections on their development and accomplishments in light of their personal learning goals. Plans for preparing the portfolio begin early in high school, and preparation of the work constitutes one important part of the curriculum. Students complete these projects over extended periods of time, sharing work in progress with teachers and peers for critique, and preparing a final exhibit of their work. The portfolio or parts of it are presented to a committee, typically consisting of the student's advisor, other teachers, and sometimes other students or members of the community. Parents may be invited to sit in on these oral defenses. The committee discusses the work with the student and then debates its merits in light of school and state curriculum standards. The decision, rationale, and suggestions are shared with the students and their parents. Students have multiple opportunities to rework and resubmit their portfolios to the committee for review. Students who disagree with the decisions have the right to appeal. Periodically, committees of outsiders, sometimes called "critical friends" (Darling-Hammond, 1994a, p. 505), engage in a quality review process, visiting classes, talking with students and teachers, examining instructional materials, auditing samples of evaluated portfolios, and raising questions to facilitate further inquiry and growth.

When we contrast this contextualized approach with a more centralized and standardized approach to certification for graduation--on demand assessments, centrally developed, and scored--the potential differential consequences are highlighted. Consider the two approaches in terms of the model of intellectual work that they present for students and teachers and in terms of the nature of discourse about educational reform that they are likely to promote.

In the typical centralized approach to assessment, all students must take the same state-administered test, after having experienced a curriculum designed to cover the capabilities assessed by the test; whereas, in the more contextualized approaches, each student develops projects, in consultation with faculty, that both suit their own interests and show evidence of having met school and state standards. This contextualized approach represents a different perspective on fairness and on the authority allocated to students and teachers in making assessment decisions: in one case allowing students, in negotiation with their teachers, to choose the products that best represent their strengths and interests and, in the other case, presenting all students with the same task after ensuring, to the extent possible, that they have had to opportunity to learn the necessary skills and knowledge.

In centralized approaches, the assessments are typically scored anonymously by readers from outside the school, and interpretations of students' capabilities are returned to the school preconstructed. Teachers, students, and parents become consumers of interpretations constructed by others. In the contextualized approach, the assessments are evaluated by teachers who know the students' work well, in dialogue with one another and with those who know the students less well, debating the merits of the performance in terms of school and state curriculum standards. Teachers engage in critical dialogue with one another, and with students and parents, to construct, critique, and revise interpretations about students' capabilities based on available evidence. State and district authorities assume more of an auditing role in certification decisions, reviewing samples of scored portfolios and procedures in place at the school level to ensure that standards of quality and fairness are being met and assisting local educators in evaluating their own programs. (See, for example, Adams & Burgess, 1992; Darling-Hammond, 1994a.) Again, the approaches reflect a different view of fairness to students and of authority allocated to teachers and students: one based on anonymity and multiple independent readings, the other based on in-depth knowledge and critical dialogue.

Case study research suggests substantial impact on the nature of teaching, learning, and community life in schools when these more interpretive approaches to assessment are introduced, including more collaboration among faculty, more public discussion and critique of standards, and, in many cases, more engagement by students in their own learning (e.g., Ancess, Darling-Hammond, & Einbender, 1993; Davis & Felknow, 1994; Mabry, 1992). Conversely, studies of reform driven by centralized approaches to assessment point to potential problems when schools are not given resources to enable teachers to reconsider their practices and try out alternatives in a supportive environment (e.g., Cohen & Ball, 1990; Corbett & Wilson, 1991; Ellwein, Glass, & Smith, 1988; Smith et al., 1994). Case studies such as these, examining the impact of different assessments on both educational communities and individual stakeholders (including those developing, taking, scoring, and using the assessment information), would aid policymakers in understanding the consequences of their choices.

Readers will note that with this approach, the locus of decision making about individual students is at the school level. In making these arguments, I am not intending to devalue information of the sort the

National Assessment of Educational Progress (NAEP) or state-level assessments like NAEP provide for policymakers. Rather, I am raising questions about whether the same test can or should be used for multiple purposes, as some states intend. Here, it is important to consider which interpretations and decisions are best made at the local level and which are appropriately controlled from a distance, and to consider what the consequences of these different choices are for the nature of teaching, learning, and discourse about educational reform.

Darling-Hammond (1994b) describes the policy in states where assessments used for policy purposes are distinct from those used for consequential decisions about individual students:

> These states envision carefully targeted state assessments at a few key developmental points that will provide data for informing policymakers about program successes and needs, areas where assistance and investment are needed, and assessment models for local schools. Meanwhile, locally implemented assessment systems--including portfolios, projects, performance tasks, and structured teachers' observations of learning--will provide the multiple forms of evidence about student learning needed to make sound judgments about instruction. (p. 20)

Darling-Hammond (1994a) contrasts this type of policy with a more centralized approach to assessment in terms of the ways in which each engenders educational reform:

> At the policy level, the top-down, content-driven approach to standards and the bottom-up approach driven by investments in teacher and school learning represent different theories of organizational change. One view seeks to induce change through prescriptions for practice buttressed by extrinsic rewards and sanctions for schools and students, on the assumption that the fundamental problem is a lack of will to change on the part of educators. The other view assumes that the fundamental problem is a lack of knowledge about the possibilities for teaching and learning, combined with an undeveloped organizational capacity for ongoing change and improvement. This view seeks to induce change by building knowledge among school practitioners and parents about alternative methods, by stimulating organizational rethinking through opportunities to work together on the design of teaching and schooling, and by providing the means to try new approaches. (p. 507)

Readers interested in additional information about validity research in the development and evaluation of interpretive approaches to assessment might consult Hipps (1992), Johnston (1989, 1992), Moss (1994), and Moss et al. (1992). In addition, methodological advice from interpretive or qualitative researchers can be profitably adapted to the assessment context to assist both in designing/evaluating assessments and in examining the consequences of using assessment (see, e.g., Denzin & Lincoln, 1994; Erickson, 1986; Guba & Lincoln, 1989; Packer & Addison, 1989). Scholars who specifically address issues of fairness and authority in critical dialogue include Guba and Lincoln (1989), Lather (1986), and Heron (1988). In addition, Smith (1993) and Moss (in progress) provide overviews of differing sets of validity criteria.

CONCLUDING COMMENTS

Perhaps the most striking contrast between interpretive and psychometric approaches is that with the psychometric approach, readers rarely, if ever, review the entire set of performances for a candidate in reaching or evaluating a certification decision (or setting standards). No one is asked to carefully read over all the evidence available on a candidate and say, "Yes, I think this is a highly accomplished candidate, and here's why."

This is, of course, one way of maintaining the objectivity and fairness of the scoring and of ensuring that the decision is not based upon premature judgments or biases of individual readers. However, with the interpretive approach, such bias can be challenged and controlled through debate, audit, and appeal; just as it is in the law. And, much information is lost when performances are abstracted in scores and certification decisions are then based upon algorithms for combining the scores and comparing them to a predetermined cut score. Why is it easier for us to accept judgments based on decontextualized descriptions or graphic profiles than it is to accept judgments based on complete cases? If we step, for a moment, outside the traditional assumptions for our profession, doesn't that seem a bit odd? It seems essential to ask whether expert judgment by highly accomplished readers, who have become familiar with the candidate's achievements across multiple performances, may not result in a more valid and fair decision.

So, I think it important to ask questions that challenge our standard practices and highlight the potential consequences of the choices we are making. Is it more valid to have exercises evaluated independently, decontextualized from the rest of the performances? Is it more valid to preclude debate among readers about the actual performances in question? I do not believe we really know the answers to those questions. We have a long history that implies the answer is yes. And yet, since we have taken the practices for granted, we have rarely, if ever, questioned those assumptions conceptually or empirically. There are traditions with equally long histories that suggest the answer to these questions is no. We have yet to take Messick's (1989) advice about the importance of comparing different inquiring systems to illuminate the assumptions and values that underlie each.

Implicit in any assessment process is a value-laden vision of what education is and of what appropriate roles are for teachers, students, and other members of the community. Comparisons among alternative approaches to assessment of the sort I have proposed, highlight taken-for-granted practices and make explicit the values that underlie them. In this way, policy debate regarding the purpose and practice of assessment becomes better informed. The research agenda described in this paper suggests one step in that direction.

References

Adams, E., & Burgess, T. (1992). Recognizing achievement. In H. Berlak, F. Newmann, E. Adams, D. A. Archbald, T. Burgess, J. Raven, & T. A. Romberg (Eds.), *Toward a new science of educational testing and assessment* (pp. 117-137). Albany: State University of New York Press.

Ancess, J., Darling-Hammond, L., & Einbender, L. (1993). The development of authentic assessment at Central Park East Secondary School. In L. Darling-Hammond, J. Snyder, J. Ancess, L. Einbender, A. L. Goodwin, & M. B. Macdonald (Eds.), *Creating learner-centered accountability*. New York: Columbia University, Teachers College, National Center for Restructuring Education, Schools, and Teaching.

Cohen, D. K., & Ball, D. L. (1990). Relations between policy and practice: A commentary. *Educational Evaluation and Policy Analysis, 12*(3), 331-339.

Corbett, H. D., & Wilson, B. L. (1991). *Testing, reform, and rebellion*. Norwood, NJ: Ablex Publishing.

Darling-Hammond, L. (1994a). National standards and assessments: Will they improve education? *American Journal of Education, 102*(4), 478-510.

Darling-Hammond, L. (1994b). Performance-based assessment and educational equity. *Harvard Educational Review, 64*(1), 5-30.

Davis, A., & Felknow, C. (1994, April). *Graduation by exhibition: The effects of high stakes portfolio assessment on curriculum and instruction in one high school.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Delandshere, G., & Petrosky, A. R. (1992). *Developing a high inference scheme to evaluate teacher knowledge.* Paper presented at the annual meeting of the Educational Research Association, San Francisco.

Delandshere, G., & Petrosky, A. R. (1994). Capturing teachers' knowledge: Performance assessment a) and post-structuralist epistemology b) from a post-structuralist perspective c) and post-structuralism d) none of the above. *Educational Researcher, 23*(5), 11-18.

Denzin, N. K., & Lincoln, Y. S. (Eds.). (1994). *Handbook of qualitative research*. Thousand Oaks, CA: Sage.

Ellwein, M. C., Glass, G. V., & Smith, M. L. (1988). Standards of competence: Propositions on the nature of testing reforms. *Educational Researcher, 17*(8), 4-9.

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 119-161). New York: Macmillan.

Guba, E., & Lincoln, Y. (1989). *Fourth generation evaluation*. Newbury Park: Sage.

Heron, J. (1988). Validity in co-operative inquiry. In P. Reason (Ed.), *Human inquiry in action*. London: Kogan Page.

Hipps, J. A. (1992). *New frameworks for judging alternative assessments.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Jacobson, L., & Stilley, L. R. (1992). *Developing and scoring assessment center exercises.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Johnston, P. H. (1989). Constructive evaluation and the improvement of teaching and learning. *Teachers College Record, 90*(4), 509-528.

Johnston, P. H. (1992). *Constructive evaluation of literate activity.* New York: Longman.

Lather, P. (1986). Research as praxis. *Harvard Educational Review, 56*(3), 257-277.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry.* Beverly Hills, CA: Sage.

Mabry, L. (1992). *Alternative assessment in an American high school.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

McDonald, J. P., Smith, S., Turner, D., Finney, M., & Barton, E. (Eds.). (1993). *Graduation by exhibit.* Alexandria, VA: Association for Supervision and Curriculum Development.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5-12.

Moss, P. A. (in progress). Shifting conceptions of validity in educational measurement II: Criteria from alternative research traditions. Manuscript in preparation, University of Michigan.

Moss, P. A., Beck, J. S., Ebbs, C., Herter, R., Matson, B., Muchmore, J., Steele, D., and Taylor, C. (1992). Portfolios, accountability, and an interpretive approach to validity. *Educational Measurement: Issues and Practice, 3*(11), 12-21.

Packer, M. J., & Addison, R. B. (Eds.). (1989). *Entering the circle: Hermeneutic investigation in psychology.* Albany: State University of New York Press.

Smith, J. K. (1993). *After the demise of empiricism: The problem of judging social and educational inquiry.* Norwood, NJ: Ablex Publishing.

Smith, M. L., Noble, A. J., Cabay, M., Heinecke, W., Junker, M. S., & Saffron, Y. (1994). *What happens when the test mandate changes? Results of a multiple case study* (CSE Technical Report 380). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Wineburg, S. (1988). *A candidate-centered approach to the assessment of teaching* (Teacher Assessment Project, Technical Report H-14). Stanford, CA: Stanford University.

An Interpretative Approach to Setting and Applying Standards

Summary of Break-out Session[1]

At the beginning of this session, participants were encouraged to present any specific concerns that they might have related to the conference as a whole, to the session, and/or to Moss' session. These concerns were listed, and they guided the discussion.

One participant addressed the issue of general and specific applications of the various approaches to setting and applying standards. The concern was that there are general applications that can be addressed by each approach, and there are, as well, some very specific applications for which one particular approach is most appropriate. The participant pointed out that in judging what the National Assessment Governing Board (NAGB) was attempting in the National Assessment of Educational Progress (NAEP)--a very specific application with tests in various subject areas and at several levels--it would be beneficial to move from a general discussion of the benefits of the method to a discussion of the particular application. Another important issue is how the methods, or any method, can contribute to NAEP's standards. Moss added that an additional concern would be how to present the results.

Another participant focused on content standards and assessment. He observed that content standards define test content. He pointed out that higher and higher content standards are being set, however, and, in measuring them, the assessment test has begun to move off center; eventually, the content will be outside the student's reach. The questions then was: What does that mean in terms of the interpretability of standards? Another participant indicated that it is important to have a statement of purpose for NAEP. She asked how the information from NAEP could be used in light of information on validity.

The following question was presented: Is the consensus building that goes into the framework of NAEP and that is used in the standards-setting process an attempt to disseminate and formulate a conceptual definition of curriculum standards? If so, then issues related to equity and standard setting are quite different from how they are being perceived at this time. One response was that the discussion needs to be extended to consensus in a pluralistic society, particularly to consensus distribution or the range of disagreement around the consensus point. The participant explained that it matters if that disagreement is caused by random error or whether it represents completely different value. If it represents value perspectives, then the average may not represent any person at all. He suggested that another question would be how one can deal with pluralism and consensus at the same time.

One participant asked if it is possible to link performance levels aggregate scores on tests to the different notion of performance standards, that is, a scoring rubric for a particular type of item, or what to look for in an excellent, fourth-grade essay. She noted that NAGB's initial goal seemed to be to find that link: a measure of aggregate performance that could be linked with the characteristics of a good essay.

---

[1] This is a summary of the break-out discussions for Pamela A. Moss's presentation. The session was facilitated by Paul Williams (Educational Testing Service) and recorded by Sheida White (National Center for Education Statistics).

BEST COPY AVAILABLE

Another participant, referring to the various ways in which Moss used the term "construct validity," asked for an explanation of construct validity and how the use of the integrative approach (described in her presentation) could assist with the understanding of standards. Moss indicated that her use of the term referred to interpretations and the consequences of those interpretations.

The group reviewed the salient points of both the psychometric approach to, and interpretative assessment process of, setting and applying standards, as well as the underlying vulnerabilities of each approach. Participants sought clarifications of their understandings of the interpretative approach. Participants pointed to the benefits of making information related to the interpretative assessment process, including a justification for its use, accessible and available to the wide audience of persons with interest in large-scale assessment.

The group explored which of the two methods, psychometric or interpretative, is the more valid in high-stakes assessment. It was decided that there is no right model to do this since each approach to assessment is characterized uniquely in terms of, first, how each treats the relationship between the parts of an assessment and the whole, and, second, how human judgment is used in arriving at a well-warranted conclusion.

One participant asked if the interpretative approach could be incorporated in a large-scale assessment like NAEP. It was suggested that we should encourage the dialogue on this issue that occurs at the national level to take place at the state, community, and parent levels.

The participants then examined the group dynamic of forming judgments when using the interpretative process. Moss delineated the process of forming judgments when using the interpretative process and indicated that there are problems in the group dynamics where individuals' voices can get lost in the presence of other dynamic members. However, she suggested, first, that each judge take personal responsibility for participation in the dialogue, and, second, that quiet members of the group be given an advocate. Moss also emphasized that the consequences of a particular approach on a system is what is important, not just the quality of information.

Some participants noted that some teachers make good judges, and they can be trained to improve their own assessments as well as to participate in the standard-setting process. The group agreed that for the sake of efficiency, consistency, and validity in the scoring of performance assessment and in setting standards, it is necessary to include people from a variety of backgrounds in the standard-setting process. Additionally, it was agreed that a system that encourages every individual to come to an evidence-based interpretation is needed.

The discussion then shifted to different types of assessments: *national assessment, school-based assessment/teacher-based assessment, and standard setting*. It was emphasized that different assessment programs have different purposes and that *school-based/teacher-based assessment* is very context-specific. The group was reminded that the goals of *national assessment and school-based assessment* are compatible; a national system monitors broad performance. Some participants pointed out that it is difficult to understand the purposes of national assessment. The group, however, agreed that no one assessment system can satisfy all the needs. It was suggested that one way of bringing national standards to the school level is to engage in dialogues on how achievement levels in the national examinations compare with those levels set by classroom teachers.

The group then focused its discussion on the issue of *opportunity to learn*. Some participants emphasized that assessment must be based on the content to which a student has been exposed. It was suggested that content standards are the blueprint for what students learn, and that the opportunity to learn standards guarantee that there is a link between what is supposed to be learned and what is actually learned.

The discussion turned to *the relationship between content standards and test specifications*. Participants asked broad questions about the relationship between content standards, test specifications, and the examination itself. Other issues discussed pertained to tests that do not fully represent the scope of the domain tested. It was suggested that these problems are not insurmountable. The group then discussed whether there should be standards set for multiple perspectives and different interest groups or whether there should be common standards. Opinions were varied.

There was a brief discussion on the concept of validity. The group was cautioned to distinguish clearly between *the concept of validity of the standards and the concept of the validity of the assessment* itself. One participant suggested that there is a need for a glossary of terms related to the new vocabulary of achievement and standards.

The Consensus Process in Standards Development

Belinda L. Collins

Director, Office of Standards Services,
National Institute of Standards and Technology

ABSTRACT

The consensus process in the development of standards is considered to include balance, openness, agreement, and due process for all affected parties. The American National Standards Institute (ANSI), all major standards developers in the United States, the National Institute of Standards and Technology (NIST), and the International Organization for Standardization (ISO) have each defined procedures for developing consensus. While they differ slightly from one another, these procedures typically concentrate on elements such as committee membership--including balance among interest groups without domination by any single entity--announcement of and openness during meetings, voting procedures, review processes, communications, and appeals. Although these procedures are generally applied to the development of product and process standards to ensure quality, safety, health, and/or environmental integrity, they may provide a useful model for evaluating the effectiveness of procedures and standards developed in the field of education. The general consensus process is described here, with particular focus on relevance for developing standards for education.

INTRODUCTION

In the United States, standards are primarily produced through a private, voluntary system that is industry led, with active participation by government. The standards process is intended to be a fair and open process in which all views can be heard as a standard is developed. The concept of consensus in standards development is considered here to represent a process in which there is balance, openness, agreement, and due process for all affected parties. The American National Standards Institute (ANSI), all major standards developing organizations such as ASTM (formerly the American Society for Testing and Materials) and Underwriters Laboratories (UL), the National Institute of Standards and Technology (NIST), and the International Organization for Standardization (ISO) have defined procedures for developing consensus. The sense of these procedures will be discussed in detail, along with the goals of the processes they use. Their relevance to the educational community will be discussed here, as well as recent research that relates procedures used by the voluntary standards community to those used by the educational community.

BACKGROUND

What is meant by "standards process"? The term "standards" covers not only the traditional product standards, but also includes system and process standards. These standards are intended to ensure public health, safety, and protection of the environment, as well as to provide performance measures to meet the challenges of global competitiveness. There are standards for products such as toasters and computers, as well as standards for processes such as the ISO 9000 series of Quality Management and Quality Assurance Standards. The term standards process refers not only to the process of writing standards, but also to the development of the test methods, certification processes, and auditing procedures that enable users of the standards--namely, manufacturers, private sector and government buyers, or independent third parties--to assess product performance. These latter activities are commonly referred to as "conformity assessment." This system addresses issues related to both product and process standards, as well as the means by which acceptable demonstrations of compliance with specified standards or regulations are generated. Both product and process standards are intended to facilitate commerce, safeguard the public's health, welfare, and safety, and protect the environment.

In the United States, the standards process is primarily private sector and voluntary, industry led and government supported. Briefly, the three major players in the U.S. standards process are industry, private standards organizations (including ANSI), and government. Voluntary consensus standards cover the gamut of industrial interests, including product specification and materials testing, electrical and fire safety, and nuclear plant construction and operation (to name only a few). These standards, which are set for test methods and terminology, product specifications, and general management practices, are produced through a consensus process in which all interested parties participate and have equal voices. Technical experts from all walks of life, including industry and government, come together to support the activities of a large number of private standards-developing organizations (SDOs; more than 600, at last count). Their efforts result in tens of thousands of consensus standards in the United States. ASTM, for example, currently has about 9,000 standards on its books covering six types of standards: test methods, specifications, practices, terminology, guides, and classifications.

The role of each of the three major players in the process is different but critical to an understanding of the complexity of the standards process. Thus, industry manufactures products to the standards that its representatives help to write as participants in SDOs. With these standards, manufacturers can sell their products and services in both domestic and international markets. Typically, the standards organizations provide secretariat services and produce documents for sale, with ANSI serving as an "umbrella" organization, as the member body of ISO, and supporting the U.S. National Committee to the International Electrotechnical Commission (IEC). Finally, the federal government develops its own regulations (sometimes using voluntary standards as input), sets some standards, procures products according to standards, and provides technical expertise to a host of standards committees.

To guide the federal government, the Office of Management and Budget (OMB) issued Circular A-119 (most recently revised in October 1993), which establishes policy to be followed by executive agencies for working with voluntary standards. This circular defines a standard as follows:

> Standard means a prescribed set of rules, conditions, requirements concerned with the definition of terms; classification of components; delineation of procedures, specification of dimensions, materials, performance, design or materials, products systems, services, or practices; or descriptions of fit and measurement of size. (p. 2)

OMB (1993) states further that "It is the policy of the Federal Government in its procurement and regulatory activities to: a) Rely on voluntary standards, both domestic and international, whenever feasible and consistent with law and regulation pursuant to law" (p. 2). It is also federal policy that the government "participate in voluntary standards bodies when such participation is in the public interest" (p. 3). OMB points out that agencies should:

> recognize the positive contribution of standards development and related activities. When properly conducted, standards development can increase productivity and efficiency in Government and industry, expand opportunities for international trade,

> conserve resources, improve health and safety, and promote the concepts of pollution prevention and the use of recycled content materials. (p. 3)

No single federal agency has oversight over the entire voluntary standards process or even the governmental role within it, however, just as no single SDO coordinates the work of the private sector. As it currently exists, the system's numerous participants have different perspectives and objectives, with no clearly defined coordination by either the federal government sector or the private sector.

Nonetheless, among federal agencies, NIST plays a unique role in support of the U.S. standards process. NIST's mission is to promote U.S. economic growth by working with industry to develop and apply technology, measurements, and standards. As part of its core mission, NIST produces the fundamental physical standards, measurements, test methods, reference data, and materials that provide the technical underpinning for standards and conformity assessment. NIST staff work actively in the voluntary standard process, holding over 1,100 memberships on more than 800 standards committees, including information technology and fire safety. Through programs administered by the Office of Standards Services, NIST accredits qualifying laboratories to make measurements needed by industry (including calibration procedures, testing for asbestos, flammable fabrics, and construction materials) and supports international trade efforts such as the General Agreement on Tariffs and Trade (GATT) and the North American Free Trade Agreement (NAFTA), as well as bilateral negotiations for mutual recognition of product approvals with major trading partners. NIST also chairs the Interagency Committee on Standards Policy (ICSP), which facilitates cooperation among federal agencies in the voluntary standards process. In all these activities, NIST is committed to serving as a productive partner for improving communication and cooperation among government and standards organizations.

THE PROCESS

In the U.S. voluntary standards system, standards are typically developed by standards-developing organizations made up of players from both the private sector and the public sector. The system is termed "voluntary" because participation is on a voluntary basis--with funding provided by a participant's employer--not by the government. Furthermore, the standards produced are intended for voluntary use. Only if they are adopted by governmental authorities do they become regulations.

Not all voluntary standards are consensus standards. *Industry* standards are developed by a particular company or industry group and are controlled by that industry. While they reflect primarily the views of the particular company, they often include input from more than one company in a particular industry group or sector. Industry standards are typically used in-house, either by a single company or by members of a trade association and often cover product specifications and design. Frequently, industry standards are intended to ensure that parts of a product will be assembled properly in the final product.

On the other hand, the *consensus* standard is typically developed by a standards-developing organization or other consensus body. It often extends well beyond product specifications to more general test methods and process definitions. Many SDOs are devoted to standards writing and derive most of their livelihood from the sale of their standards. Representative SDOs in the United States include ASTM, UL, and the National Fire Protection Association (NFPA), to name only a few. As an important element of their activities, professional societies, such as the American Society for Mechanical Engineers (ASME) and the Institute of Electrical and Electronics Engineers (IEEE), as well as industry trade associations, such as the National Electrical Manufacturers Association (NEMA), also develop consensus standards according to rigorously defined procedures. In all cases, the term consensus means broad participation by a diverse group of affected parties most of whom are typically members of the organization. Since a consensus standard is intended to reflect the view of all interested and affected parties, other interested constituencies such as academicians and consumers are generally welcome to participate in the standard's development.

Most consensus standards are developed through committees that have been duly constituted for the purpose of developing and revising standards. These committees work to develop standards in specific domains and are generally composed of producers, users, and "general interest" members. Typical participants in the consensus standards-developing process include technical professionals, academics, consumers, labor representatives, government at all levels, industry, and legal professionals who all come together in a voting committee under the aegis of the SDO. The SDO is responsible for ensuring due process, as well as for maintaining all records of the development process, and usually for publishing the final standard. Participation in the committee must not be dominated by any one group and, in fact, must include at least three of the typical participants listed above. The committee process, however, is controlled by the procedures of the SDO (rather than by a particular industry or entity).

PROCEDURES FOR DEVELOPING VOLUNTARY STANDARDS

Most standards developers have established guidelines or procedures for developing standards. These guidelines include membership criteria, voting procedures, formats, appeal processes, and revision intervals. The rigorous procedural requirements and diverse representation within the SDO are hallmarks of the consensus process. ANSI (1993) defines consensus as follows:

> . . . substantial agreement has been reached . . . by directly and materially affected interest categories. Substantial agreement means much more than a simple majority, but not necessarily unanimity. Consensus requires that all views and objections be considered, and that an effort be made toward their resolution. (p. 3)

ASTM (1992) defines consensus as follows: "the judgment arrived at through the balloting and review procedures of these [ASTM] regulations" (p. 5). These procedures are applied at each step of the development process. While the ASTM procedures are among the most rigorous, and will be described here, many SDOs have similar procedures. ANSI has also developed procedures for accrediting standards developers--based on the use of agreed-upon procedures.

The requirements for consensus procedures are replicated at the international level by both ISO and the International Electrotechnical Commission (IEC). Each of these entities uses consensus and due process as key elements of their standards development process. ISO/IEC (1992) defines consensus in the following manner:

> General agreement, characterized by the absence of sustained opposition to substantial issues by any important part of the concerned interests and by a process that involves seeking to take into account the views of all parties concerned and to reconcile any conflicting arguments. (p. 28).

ISO emphasizes this statement with a note that "consensus need not imply unanimity" (p. 28).

DUE PROCESS

Over the years, standards organizations have adopted agreed-upon procedures for determining and developing consensus through due process. ANSI (1993) defines due process as:

> Due process means that any person (organization, company, government agency, individual, etc.) with a direct and material interest has a right to participate by: a) expressing a position and its basis, b) having that position considered, and, c) appealing if adversely affected. Due process allows for equity and fair play. (p. 1)

In the following section, the ANSI procedures are used as an example, although most standards-developing organizations have somewhat similar procedures. The overriding principles of due process in supporting the consensus standards process are those of openness and balance, with participants drawn from different key interest groups (so that no single interest dominates the standards-development process). Due process requires written procedures for the notification of proposed standard development, the development process, an appeals mechanism, consideration of all views and objections, consideration of proposed standards action, and complete record keeping. Gross (1994) defined openness as meaning that participation shall be open to all persons who are directly and materially affected by the activity in question, with timely and adequate information provided on the initiation of development of a new standard or revision of an existing standard. Balance-of-interest categories is crucial so that no single party can dominate the process. ANSI (1993) further specifies that at least three interest categories must be represented in the standards development--namely, producer, user, and general interest--although more categories may be added as appropriate to the nature of the standard. The written procedures for developing the standard must be available to all interested parties and must include "an identifiable, realistic, and readily available appeals mechanism for the impartial handling of substantive and procedural complaints" (p. 2). ANSI further provides for announcement of standards activities in appropriate media (such as the *ANSI Reporter*) so that all interested parties are informed of the opportunity to participate. Another key element of due process is the requirement to "give prompt consideration to the written views and objections of all participants" (ANSI, p. 2). Prompt attention must also be given to proposals for withdrawing, revising, or developing new standards. Finally, all records of the development process must be maintained to demonstrate compliance with the written procedures. These records are essential for determining if due process was followed should a third party challenge the standard's development.

The ANSI (1993) due process requirements can be summarized as follows: openness, balance, interest categories, written procedures, appeals, notification of standards development, consideration of views and objections, consideration of standards proposals, and records. There are five key principles within these requirements for ensuring equity, as follows:

1. Adequate notice of proposed actions (development, withdrawal or revision),
2. Ample provision of opportunities for participation,
3. Adequate records of all discussions and decisions by the participants,
4. Timely distribution of minutes and ballot results, and
5. Careful attention to minority opinions.

Some standards developers also submit their standards to ANSI for recognition as American National Standards (ANS). These standards are often issued as joint ANSI/individual organization standards. ANSI recognition of these developers means that they must provide evidence that their standards have been produced according to consensus and due process, meaning that any person, organization, company, government agency or individual with a direct and material interest has a right to participate. Due process allows for equity and fair play (Gross, 1994).

ANSI (1993) accredits standards developers to use one of three different methods for developing consensus. These are the Accredited Organization Method, the Accredited Standards Committee Method, and the Accredited Canvass Method. Each method specifies procedures to be used during the development of a standard to ensure that all agreements truly reflect consensus.

According to ANSI (1993), the Accredited Organization Method is used by many societies that develop standards. While participation is open to any and all interested parties, members of the consensus body are often members of the society. This method requires the standards developer to develop its own operating procedures, which must meet the general requirements of the ANSI procedures. This process allows the standards developer to create and use a system that suits its own practices but that is consistent with the ANSI procedures.

The Accredited Standards Committee Method (ANSI, 1993) is used by

> standing committees of directly and materially affected interests created for the purpose of developing a document and establishing consensus in support of this document for submittal to ANSI. The Committee Method is most often used when a standard affects a broad range of diverse interests or where multiple Associations or Societies with similar interests exists [sic]. The committee serves as a forum where many different interests, without a common membership in an organization or society, can be represented. (p. 30)

A secretariat administers these committees and is responsible for providing administrative oversight and ensuring compliance with relevant procedures. These committees may adopt the ANSI standard procedures or develop their own procedures based on the ANSI requirements.

The Accredited Canvass Method (ANSI, 1993) is most frequently used by "smaller Trade Associations or Societies that have documented industry practices and wish to have these standards recognized nationally. Most canvass developers are responsible for less than five standards each" (p. 30). In this process, the SDO identifies people who are likely to be "*directly and materially affected*" (p. 30) (including notification in *Standards Action*). The SDO then conducts a letter ballot (or canvass) of these people to determine their positions on the document. In this method, due process begins *after* the draft standard has been developed. ANSI requires that SDOs using this method apply ANSI procedures to ensure consensus and due process.

Underwriters Laboratories (UL) uses both the Canvass Method and the Accredited Organization Method in developing its standards. Consensus is required in both methods. Following ANSI's lead, UL (1992) defined consensus as "substantial agreement reached by concerned interests according to the judgment of a duly appointed authority, after a concerted attempt at resolving objections. Consensus implies much more than the concept of a simple majority but not necessarily unanimity" (p. 6).

PROCEDURES FOR DETERMINING CONSENSUS

While most standards developers have explicit procedures for determining consensus, the ASTM procedures (1992) are particularly detailed and are used here as an example of procedures that are effective in supporting the consensus process. ASTM provides specific procedures for balloting items by letter ballot (written vote) at the subcommittee level, then the main committee level, and finally the society level. At each level, specific percentages of affirmative votes are required. Thus, at the subcommittee level, an affirmative vote is defined as "two thirds of the combined affirmative and negative votes cast by voting members with not less than 60% of the voting members returning letter ballots" (p. 11), while at the full committee level, the percentage of affirmative votes required increases to 90% of the total votes received. A negative vote must be accompanied with a written statement explaining the reason for the vote. This procedure lessens the likelihood of objections based on personalities or side issues and focuses the discussion on technical, substantive issues. Negative votes must be acted on at either a subcommittee meeting or by a letter ballot--with the reasons for the negatives circulated to all committee members, preferably in advance of the meeting. A subcommittee's recommendation that a negative vote is not persuasive or not related to the item being balloted is accepted only by an affirmative vote of at least two thirds of those voting. The rules for considering negatives apply as a standard progresses from subcommittee to main committee and then to the society. Resolution of negatives is done primarily by the subcommittee; the main committee typically refers negatives back to the subcommittee for further consideration. It is important to understand that there is a *process* for objecting to items contained in draft standards and a *process* for considering and resolving these objections. Use of these processes frequently results in major revisions and changes to draft standards and provides a means by which all voices can be heard.

In the case of doubt about consensus, the ISO procedures require an affirmative vote by a two-thirds majority of the full members of the committee (or subcommittee) for a draft to be considered a draft international standard. ISO states that every attempt must be made to resolve negative votes. A draft international standard is approved by a two-thirds majority vote of the member nations if not more than one quarter of the total votes cast are negative. Again, similar processes pertain among the various national bodies, although the percentages may differ slightly. At the international level, a draft may circulate for as long as six months to allow all member nations to vote and to develop a consensus among their members for that vote. At the domestic level, time periods are typically shorter--on the order of 60 to 90 days. ANSI also operates a Board of Standards Review to which appeal may be made if an affected party believes that due process was not followed during the development of a standard. The appeal is based on the failure to follow procedure, not failure to include or exclude particular technical material from the standard.

The organizations discussed so far are private sector organizations in which federal and state government officials participate but without any specific "management" role. Other organizations are comprised primarily of state government officials, with industry and the federal government providing an advisory role. One such organization, which also uses the consensus approach and brings together state regulators and officials to develop standards, is the National Conference on Weights and Measures (NCWM) established in 1905 by the then National Bureau of Standards (now NIST). NCWM is a national professional organization that develops consensus standards in such areas as weighing and measuring device regulation, commodity regulation, motor-fuel quality, and administration of regulatory weights and measures programs. NCWM "is sponsored by the Federal government to provide national uniformity in State and local laws, regulations, and methods of test" (Brickenkamp, 1995, p. 2). NIST provides the secretariat for the conference, as well as technical measurement, training, and advisory resources.

Regulatory weights and measures activities are aimed at maintaining equity in the marketplace so that businesses can compete fairly and buyers and sellers can make informed decisions in trade. It is a year-round standards development, product-testing, and training organization.

> The NCWM is a voluntary standards development membership organization. Active members must be weights and measures regulatory officials from State or local agencies. Associate members are representatives from business or trade associations. Advisory members are all other members, either Federal government representatives, educational institutional representatives, consumers or consumer groups, retired regulatory officials, foreign government representatives, or other State and local agency representatives that are not weights and measures officials. In 1994, there were approximately 3,400 members of whom 52% are associate (business) representatives. (Brickenkamp, 1995, p. 7)

Because the NCWM has proved to be an effective means for bringing state regulatory officials together to develop consensus standards of mutual interest, it could serve as a model for developing and setting standards for the educational community. In this model, the federal government provides guidance and participates as a partner in the standards-setting process but neither controls nor mandates regulations.

COMPARISON OF STANDARDS DEVELOPMENT PROCESSES

The procedures specified by ANSI and put into practice by most SDOs, including NCWM, ensure consensus in the development of standards. Use of consensus, due process, and standard procedures has resulted in the development of national and international standards that facilitate trade and ensure public safety, health, and protection of the environment. The success of these procedures suggests that they could profitably be used by the education community to facilitate development of consensus standards for education.

In fact, Walser examined the approaches used by both the voluntary private sector and educational standards organizations to provide some guidance for the education community as it struggles with developing standards. In 1989, he completed dissertation research comparing the use of procedures for developing voluntary standards in educational organizations with those used by private sector, standards-developing organizations. Walser believed that the application of standards management procedures to the educational standards process could be a fruitful means of improving standards setting for education. Walser stated that:

> The standard for educational services used by the education industry in the past clearly do [sic] not suffice for the future. Standards are conceived by most educators as static constructs--as imposed requirements. The need now is to look at standards, and ways of developing standards, that can guide and support the changes the future will bring while improving the quality of educational services. The processes for developing standards for education, as well as the standards themselves, must be flexible enough to accommodate inevitable changes. (p. 1)

Walser (1989) stated that the education industry has failed to develop professional, voluntary standards that are agreed upon by a broad consensus of affected parties throughout the nation: "Education has no forum or process to use for bringing the segments of the education industry together to address, professionally, the issue of industry-wide standards for education" (p. 17). Walser's search of the literature

> did not find reference to any organized forum in education at any level, where individual and group biases and ideas could be brought out into the open; or where all who would be affected by proposed changes could have an opportunity to become involved in developing agreements on how to deal with education problems, and where the knowledge cumulatively held by practicing professional educators could be distributed quickly and freely among local, state, national, and federal education officials. (p. 19)

These observations led Walser (1989) to point out the need for horizontal standards that cut across artificial boundaries in education as well as the need to develop management strategies. These strategies "enhance and protect the local responsibility for providing education, while ensuring that the services available at local and state levels were of the highest possible quality" (p. 2). Walser found that:

> in education, there was no coordinated, national standards development process that addressed recurring problems, responded in a systematic way to the biases and interests of all segments of the education industry, or provided information to the general public about the expectations for quality levels in education services. (p. 4)

By contrast, private industry successfully used what Walser termed a "management process" for developing standards according to due process and consensus procedures.

Comparing the dismal picture of standards development in the education community with that of industry and business, Walser (1989) observed that some 400 organizations develop voluntary standards: "These standards represent the glue that holds our economy together and provides the assurance of basic quality and interchangeability of products, systems, and services that are at work in our society" (p. 45). Central to the process used to develop voluntary standards is the idea of consensus in which disparate groups of affected parties are brought together to come to a common agreement to write a standard for a particular topic--as discussed earlier in this paper. Walser concluded that "A voluntary consensus standard is, then, both voluntary in its development and in its use, and in addition, the developers of the standard must have adhered to and complied with a documented, democratic, consensus developing management process in its formation" (p. 49). Walser noted further that membership in the standards development body must include all affected parties and not be limited solely to experts or to government.

In the research conducted as part of his dissertation, Walser (1989) compared the responses of four organizations--ANSI, ASTM, National Council for the Accreditation of Teacher Education (NCATE), and the American Association of School Administrators (AASA)--to 29 questions identifying the steps each used to develop standards, including process, components, and procedures. The results indicated that the four organizations used identical procedures on only about half of the steps. Failure to use predefined procedural steps was greatest for the education organizations, which were missing many of them.

Walser (1989) determined that key areas of concern in the response by the educational associations were as follows:

1. No process for finding or using broad, representative distribution of voting members to avoid single interest control or domination by absentee members (compared with ASTM where all concerned parties had the opportunity to provide input to a standard under development);

2. No agreement on the need to follow due process procedures, but rather a reliance on a collegial approach with ad hoc discussions and idea sharing;

3. No pre-existing set of procedures for developing standards, with reliance on feedback from only a few people or select groups with no formal procedures for obtaining consensus, or even for the process for developing a standard;

4. No defined procedure for reviewing and resolving substantive votes or for revising the standard if an argument is found to be persuasive;

5. No formal process for notifying the public of steps achieved in the standards-development process (compared to the *ANSI Reporter* that provides public notice of the intent to develop a standard, committee meetings, availability of draft standards, requests for public review, and final action on standards, processes that help ensure due process and true consensus in standards development);

6. Little formal agreement on the need to make the standard open to public review before final approval;

7. No agreement on an appeals process or procedures for challenging a standard (such as the ANSI Board of Standards Review); and

8. Lack of pre-established revision and/or expiration dates for standards (in contrast to the five years that ISO, ANSI, and ASTM set as the typical "life" of a standard).

Respondents from the educational community reported that negative comments and votes were resolved through majority votes or through persuasion--not through any formal process such as that used by ASTM. Walser (1989) stated that:

> ASTM and ANSI respondents indicated an established procedure was in place for providing assurance that single negative votes, if found persuasive, could influence the outcome of voluntary standards. NCATE and AASA used the majority rule approach, or collegial discussions for resolution and respondents indicated no written procedure was in place to address single negative votes. (p. 106)

Walser commented further that both ANSI and ASTM had written procedures in place that were, in essence, standards for developing standards.

Walser (1989) concluded that there was no accepted procedure for developing voluntary consensus standards for education, with very limited interaction with the existing private sector, voluntary standards organizations as well. Walser commented further that the education establishment was accustomed to working under laws and regulations developed at either or both the state and federal levels to which they had little possibility for formal input. Examination of his data suggests no accepted procedure for identifying and selecting participants for the standards development process either.

While Walser (1989) did not point out parallels with NCWM, discussed earlier, this organization has successfully addressed the issue of developing national standards for weights and measures by representatives from state and local jurisdictions, with participation by industry and interested federal observers. This model might be equally applicable for developing a broad national body--not a federal body--for developing and implementing consensus standards for education.

IMPLICATIONS OF THE CONSENSUS STANDARDS PROCESS FOR EDUCATION

Examination of Walser's (1989) results, combined with an understanding of the processes and procedures by which the voluntary standards community develops broad consensus standards, can provide some ideas that the education community might consider for guidance in developing its own procedures. Key to the effectiveness of these processes is the use of balanced committees representing all affected interests and using consensus and due process to develop standards. The "national conference" of affected state representatives is one means for involving participants nationally, but not federally. In this respect NCWM may prove a useful model.

Still another model is provided by ISO, which has undertaken development of process management systems for activities such as quality and environmental management systems. In 1987 ISO published a series of five international standards (9000 through 9004) which were revised in 1994, on quality systems and the use of quality management systems. The ISO 9000 standards should also be studied for their applicability to the development of horizontal, process management standards that cut across industries. They describe procedures and processes for achieving quality that can be used by any industry--rather than specific targets such as pollution levels or content knowledge. In a discussion of ISO 9000, Breitenberg (1993) stated that:

> Common elements in ISO 9001, 9002 and 9003 include the need for: an effective quality system; ensuring that measurements are valid, that measuring and testing equipment is calibrated regularly; the use of appropriate statistical techniques; having a product identification and traceability system; maintaining an adequate record keeping system; having an adequate product handling, storage, packaging and delivery system; having an adequate inspection and testing system as well as a process for dealing with nonconforming items; and ensuring adequate personnel training and experience. (p. 3)

The ISO 9000 (1994) approach includes the use of third-party registrars who register (certify) that a supplier's quality system meets the standards. This registration pertains to the system, not to any specific product produced under that system. ISO 9000 is not tailored to a specific industry; rather, it is intended to be used by any of a broad range of users to determine that the management systems they use will produce products of a consistent quality. Other quality approaches, such as the Malcolm Baldrige Award administered by NIST for the Department of Commerce, strive for continuous improvement of quality and should result in ever improving products (or zero defects).

The ISO 9000 standards (1994) are but one example of the variety of standards produced by the voluntary standards community, which has successfully developed process and procedural standards both product and content oriented. Voluntary standards range in complexity from simple component specifications to intricate, complex processes. They cover a myriad of activities, processes, and

BEST COPY AVAILABLE

products in such divergent areas as standard wattages for light bulbs, test methods for evaluating chemical purity, and standard procedures for life safety and quality management. The voluntary standards community has successfully assembled sets of effective standards from many different entities, thus allowing test methods to be separate from performance criteria or content specification --and all within the framework of a common topic such as quality or life safety. The consensus approach for the diverse range of elements covered by private, voluntary standards might serve as a useful model for the education community, which has often tended to focus on content, rather than process, standards. Consideration of the models used successfully by private sector or mixed, private-public sector organizations, both nationally and internationally, could be fruitful for the education community in the United States.

CONCLUSIONS

The private, voluntary standards organizations, such as ASTM, ANSI, and ISO, have developed effective procedures for managing the development of fair and appropriate standards. These procedures provide for balance among all affected interests through the use of due process, consensus, and written procedures for developing standards. Use of these procedures could be beneficial to the education community, providing a means for bringing together all affected parties--including the employers who benefit from employing the students produced by the process--to define the needs and to develop standards for responding to these needs. Such consensus procedures have been successfully used by the voluntary standards community for many years, as well as by the federal government and the states through programs such as the NCWM and the Voluntary Product Standards Program for the softwood lumber industry (NIST, 1994). Each of these standards developers has been able to bring together balanced representation of industry, consumers, users, and experts to develop standards that meet their needs. The education community would benefit from considering these approaches as it struggles with the issues of standards.

References

American National Standards Institute (1993, September). *Procedures for the development and coordination of American national standards*. New York: Author.

American Society for Testing and Materials. (1992, July). *Regulations governing ASTM technical committees*. Philadelphia: Author.

Breitenberg, M. (1993, April). *Questions and answers on quality, the ISO 9000 standard series, quality system registration, and related issues* (NISTIR 4721). National Institute of Standards and Technology.

Brickenkamp, C. (1995). *Weights and measures in the United States* (pp 1-16).

Gross, J. (1994, October). *Presentation on the consensus process*. National Institute of Standards and Technology.

International Organization for Standardization. (1994). *Standards 9000 - 9004*. Geneva: Author.

International Organization for Standardization/International Electrotechnical Commission. *Directives: Part 1 - Procedures for the technical work* (2nd ed.). Geneva: Author.

Office of Management and Budget (1993, October 20). *Federal participation in the development and use of voluntary standards* (Circular A--119 revised). Washington, DC: Author.

National Institute of Standards and Technology (1994, March). *American softwood lumber standard* (Voluntary Product Standard PS 20-94). Author.

Underwriters Laboratories (1992). *Method of development, revision and implementation of UL standards for safety*. Northbrook, IL: Author.

Walser, F. L. (1989). *Similarities and differences in procedures for developing and approving voluntary standards in selected organizations in education and the private sector*. Unpublished doctoral dissertation, Brigham Young University, Department of Educational Leadership.

The Consensus Process in Standards Development

Summary of Break-out Session[1]

The discussion centered around six issues: (a) committees, (b) the process of setting standards, (c) characteristics of standards, (d) measurement issues, (e) consequences or outcomes, and (f) sorting out the roles of the agencies involved with the National Assessment of Educational Progress (NAEP).

The group acknowledged that *committees should be balanced*, that is, representing interests of all affected sectors, and that this balance must be evident from the beginning of the process. The roles of the committee chair are to ensure balance and to include the relevant experts on the committee. The group recommended that the committee be either appointed or self-selected.

The group then considered the *type of members* that should be considered. One participant suggested that in trades (e.g., carpentry) committee members usually include representatives from government and industry and from both producers and users. It was explained that users are important because they ensure that content reflects current practice. Another participant, following this argument, suggested that since in educational settings the end-user population is the entire population, representatives from the entire population need to be involved in the process of standard setting. The group further agreed that the committees should be composed of a broad range of experts, including subject-matter specialists, generalists, and persons who can identify the skills and knowledge needed to transfer to the workplace.

The discussion then shifted to *the process of setting standards*. The group applauded the shift from the "top-down," hierarchical decision making process used thus far to the present participatory approach. It was noted that in this latter approach, the process must be an open one with adequate opportunity for every member to be heard. It was suggested that if the prerequisite of a balanced committee is met, standards should be set by the principle of one person-one vote as follows:

1. A draft of the standards is generated, followed by comments from committee members.
2. If a member disagrees with the proposed standards and this disagreement is related to a technical issue, the committee resolves the issue based on the best technical argument.
3. The draft is revised and resubmitted for comments.

The group examined the tensions and difficulties of this process, for example, how the committee decides on when to cut off debate. One method is for the committee to vote when to stop debate, assuming that all have had an opportunity to be heard. One participant observed that while openness is essential to achieving consensus, it is necessary for technical standards to be sufficiently rigorous. Some additional information was provided about the process of the process of setting standards:

---

[1] This is a summary of the break-out discussions for Belinda Collins' presentation. The session was facilitated by Steve Gorman (National Center for Education Statistics) and recorded by Mary Naifeh (National Center for Education Statistics).

1. Decisions on standards do not have to be unanimous; there are criteria for accepting a standard.

2. The process is a slow one.

3. A problem can occur when some participants are not heavily invested in the process.

4. Setting standards for education is a political process; broad-based committees are necessary.

The group acknowledged that in the standard-setting process, if the committee is balanced, members are trained, the rules are followed, meetings are held as required, and voting procedures are appropriate, then the committee members should be able to deliver useful products.

The discussion then focused on some of the *characteristics of standards*. Some contributions by the group were as follows:

1. The standards (or scale) must be understandable, attainable, and reasonable.

2. Consensus from a broad-based group facilitates acceptance and implementation of standards.

3. The judges' scores are elements of the standard. Others should be involved in the standard-setting process.

4. There is a need for agreement in the type of standards to be set: minimum, maximum, optimum, or adequacy of the standards.

The group addressed the issue of setting different standards for different groupings of students. The opinions were varied, and the issue was addressed from different points of view. Some participants suggested that the required level can vary across tasks, depending on task difficulty, and how critical it is. Others suggested that in setting standards, consideration ought to be given to employers and their needs. Yet others suggested that it is possible to set more than one standard where the intent is to describe the distribution of scores.

The group then considered how, when, and under what conditions standards change. It was suggested that given that setting standards is a fluid process based on task requirements and user capabilities, standards change when the task changes. One participant recommended that standards be revisited at least every five years with attention given to making scales comparable over time. It was noted that setting (and changing) standards is a political issue and that consideration also needs to be given to anticipating the impact.

The discussion then focused on *measurement and standard setting*. The group agreed that there seems to be a silent, general consensus that current procedures for standard setting are not working and that measurement is the major issue. It was suggested that this problem is complicated because the very large data-user population does not always understand the complexities of measurement, standards, or the scores. Users, especially employers, however, come in contact with students who are supposed to have met the standards.

The discussion then extended to the *consequences and outcomes of standard setting*. The group focused on two areas. First, participants indicated that setting standards can provide a method for goals setting and for determining achievement levels, but that the process may close off pathways to alternative standards. Second, the participants focused on what can be done for students who do not meet the standards. It was suggested that the answer may be in the allocation of resources.

The final part of the discussion period focused on the clarification of the *roles and relationships of different agencies: the National Center for Education Statistics (NCES) and the National Assessment Governing Board (NAGB)*. After much discussion, the group summarized as follows: *NAEP* is a program of national assessment of what students know and can do in specific content areas and grade levels. It is administered by a federal statistical agency, *NCES*. *NAGB*, an advisory board, sets standards for the National Assessment of Education Progress (NAEP).

Methodological Issues in Standard Setting for Educational Exams[1]

William A. Mehrens

Professor of Education, Michigan State University

ABSTRACT

This presentation focuses on some fairly broad but interrelated concerns to keep in mind while holding discussions regarding standard-setting methodologies. A brief review of the literature, including some recent approaches to standard setting, is presented. Points of agreement and disagreement among experts are noted. The presentation concludes with some issues that must be considered that are separable from, but related to, choosing a method and some thoughts on preparing policymakers.

This paper will first discuss briefly some broad but interrelated concerns to be considered regarding standard-setting methodologies. Next, I briefly review the literature prior to 1992, discuss some of the most recent approaches to standard setting, and list some areas of agreement and disagreement among the experts. I conclude with some issues that must be considered with regard to choosing a method and some thoughts on preparing policymakers.

CONCERNS RELATED TO DISCUSSIONS ABOUT STANDARD-SETTING METHODOLOGIES

The whole purpose of this conference is to promote understanding of issues in standard setting that range from theoretical perspectives to policy perspectives. Like many of you, I have read discussions of the issues in the literature and have been engaged in many formal and informal discussions of these issues. In many of those printed and oral discussions, there has been considerable confounding of the issues. Thus, the discussions are not always as clear and useful as would be desired. In this section, I will raise seven concerns that I think should be kept in mind when discussing the specific methodologies regarding standard setting. Keeping these in mind should allow for more useful subsequent discussions.

Separate Policy/Political Views from Technical Considerations

In my view, we should attempt to disaggregate as much as possible our policy/political views from our evaluations of the technical/scientific merits of any given standard-setting methodology and the "accuracy" of any given set of standards. Certainly both policy views and technical merits are important in any decision regarding whether to set and report standards, but they are separable.

A close analogy is the issue of consequential validity. Some recent writers have posited that the notion of validity should extend beyond the accuracy of inferences made from the scores to encompass the social consequences of testing (e.g., Messick, 1989; Shepard, 1993). While *both* accuracy of inferences and social consequences are very *important*, they are *separable*, and there is some concern that broadening the concept of validity into a consideration of social concerns will cause it to lose some of its scientific meaning (see Wiley, 1991).

---

[1] A minor portion of this manuscript was adapted from *Standard Setting for Medical Board Exams* prepared by the author for the American Board of Emergency Physicians.

At any rate, there may be positive social consequences from reporting inaccurate data (e.g., on occasion, lying results in positive consequences). Likewise, there may be negative social consequences from reporting accurate data. The consequences should be considered separately from the accuracy of the data. Have we blurred the distinction in discussions of setting standards? Some believe so. Certainly both perceived consequences and technical considerations get considered within some reports. That is acceptable as long as the writers (and readers) do not let the perceived consequences impact the perceived technical merits of the methodology.

I believe it is safe to say that there is some difference of opinion about *whether* policy views have ever impacted any technical evaluations. I *hope* that we could agree that we should *not* allow an evaluation of the *technical* merits of a methodology to be influenced by such views. Further, we should not conclude that perceived negative (or positive) consequences indicate inaccurate (or accurate) standards. If, for example, we wished to assess what the public *thinks* about what level students *should* achieve to be considered advanced in mathematics, the accuracy of determining public opinion is separable both from the consequences of reporting that perception and from the accuracy of the definition of advanced. (Of course, if we knew what level was advanced, it might be preferable to tell the public rather than to ask them--assuming they would accept our definition.)

Recognize that Context Matters in Choosing Standard-Setting Methods

Standard-setting methods should not be compared in the abstract. Some methods are better in some contexts; others are better in other contexts. This conference deals with setting standards within educational settings But even within this broad context, there are many different purposes and settings. There are educational licensure decisions, employment decisions, and teacher certification decisions. We might wish to set standards for high school diploma requirements or for determining whether fifth graders should receive special interventions. We might want to set standards when there are no implications for the individuals tested--such as setting standards for the National Assessment of Educational Progress (NAEP) or to just inform the public about the level of achievements on some state-developed test.

Obviously these contexts differ regarding the level of the stakes, the costs of false positives and false negatives, the kinds of data available to the standard setters, and so on. While a contrasting-groups approach (using criteria external to test performance) may work in setting standards for fifth-grade intervention decisions, it is commonly accepted that there is no good way to place individuals into the two groups for licensure decision making.

The context (and the specific type of assessment used) may impact whether we would prefer to employ a compensatory, conjunctive, or disjunctive model (or some combination of the compensatory and conjunctive models). The conditions under which each model may be preferable have been discussed elsewhere (e.g., Mehrens, 1990). The point is that a preference for one type of data-combination methodology impacts the choice of a standard-setting methodology. The context and specific assessment formats also may impact whether the assessment provides unidimensional data that, in turn, may impact the data-combination model chosen and the standard-setting procedure.

Recognize That Type of Item May Impact Standard-Setting Methodologies

Much of the research on standard-setting methodologies has been done using multiple-choice items. When other item formats are used, we may wish (or be forced) to change methodologies. For example, the Nedelsky (1954) method can be used only on multiple-choice items. The Angoff (1971) method has been used primarily for such items and must be modified at least slightly for nondichotomous items. If more than one item type is used on the same test, then decisions must be reached regarding whether they will all be placed on the same scale and, if so, what the scaling methodology will be and whether the standard-setting methodologies need to change across formats.

Don't Overgeneralize About "Best Methods" Across Contexts

Because of all the context effects discussed previously, we should be very cautious about generalizing across contexts. I do *not* share the conclusions the National Academy of Education (NAE) Panel has drawn regarding the Angoff approach. However, even if it were correct about Angoff in the context it investigated, it is my opinion that the report by the NAE Panel is insufficiently cautious about generalizing. The report does state that the "findings" may not "generalize to other testing situations" (Shepard et al., 1993, p. 60) and that the "Angoff procedures may or may not be defensible in other contexts" (p. 119). However the report also states that "the NAE Panel is understandably skeptical" (p. 119) about the defensibility of the procedure in other contexts.

Don't Ignore Previous Research and Draw Conclusions on a Single Study

There have been more than 20 years of research on standard-setting methodologies. That research has not been definitive, but neither has it been without merit. Although researchers may correctly believe that their research has contributed some new information to the accumulated findings, it is both immodest and unscientific to ignore past research. As Cizek (1993) points out in response to the NAE Panel's report, "the literature of standard setting provides ample documentation that the Angoff method is a reasonable, useful, acceptable, and--in many circumstances--preferable method for deriving cutting scores" (p. 5). Mullins and Green (1994) "provide our years of practical experience that have shown a modified Angoff procedure works well" and that "The Angoff technique and its variations, while not perfect, are the current standard-setting methods of choice in the certification and licensure community and offer the most stable results" (p. 23). It is particularly hazardous (as the NAE Panel has done) to "employ methods that are less well researched, and as a result less well understood, than the Angoff procedure as the basis for empirical checks on the Angoff results" (Kane, 1993, p. 3).

Separate Views About the Quality of the Item Pool from the Evaluation of the Standard-Setting Methodology

The Stufflebeam, Jaeger, and Scriven (1991) evaluation of the National Assessment Governing Board's (NAGB) efforts to set achievement levels raised questions regarding the quality and adequacy of the existing National Assessment of Educational Program item pool for setting standards on what students should be able to do. This was particularly true at the advanced level. This is certainly a relevant (but arguable) issue. However, discussion of this issue needs to be kept separate from generalizing about a procedure. I submit that if the item pool is inadequate, *no* standard-setting methodology could set an appropriate standard on the test. (I wish to be clear that I *do not* believe that Stufflebeam et al. confused the issues of item-pool quality and standard-setting *methodology*. I do believe subsequent discussions by others have confused the two issues.)

Recognize That Item Exemplars and Standard Setting are Separable Issues

One approach to communicate to the public about where the standards are set is to use item exemplars. The adequacy of this technique for communicating is problematic for a variety of reasons. There is no necessary relationship between the adequacy of these exemplars as communicators and the "correctness" of the standards.

It should be stressed that the above paragraph relates to item exemplars as communication devices to the public. If items, or more commonly, achievement level descriptors, are used to assist panels in understanding the meaning of terms such as basic and proficient, then the adequacy of these descriptors and exemplar items is of importance (see Kane, 1993, p. 5).

REVIEW OF THE PRE-1992 LITERATURE

The literature reviewed was obtained through my personal library and through computer searches using Educational Resources Information Clearinghouse (ERIC) and Psychological Literature (PSYCLIT). A very thorough review by Berk (1986) served as a major source of materials prior to 1986. Other *major* references consulted were Jaeger (1989, 1990b) and Livingston and Zieky (1982). Berk, Jaeger, Livingston, and Zieky are recognized as leading authorities on standard setting. Two other important references that provided thorough reviews were dissertations completed at Michigan State University by Korte (1987) and Cizek (1991).

General Review Findings

Although the research is not as complete or as compelling as we might wish, there is a large body of existing literature on standard setting. Although the literature on standard setting is inconclusive on many points, there does seem to be agreement that (a) setting defensible standards is difficult, (b) standard setting is a judgmental process, and (c) "a right answer does not exist, except, perhaps, in the minds of those providing judgments" (Jaeger, 1989, p. 492). There is no way to prove that any particular standard is better than any other because the correctness depends upon one's values (Zieky, 1989). The setting of standards is inherently political (Popham, 1987).

Professional *Standards* and *Guidelines*

In any review of what is an appropriate psychometric procedure it is prudent to examine what is stated in two documents: *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA/APA/NCME], 1985), hereafter called the *Standards*, and *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission [EEOC], 1978), hereafter called the *Guidelines*. (There is some debate about whether the *Guidelines* apply to many educational examinations.)

The *Standards* do not specify a particular process for setting the standards for a test. In fact, in discussing the development of the *Standards*, Linn (1984) stated that "there was not a sufficient degree of consensus on this issue . . . to justify a specific standard on cut scores" (p. 12).

Nevertheless, there are statements in the *Standards* that speak to the issue of standard setting. Standard 3.1 (1985) states that "tests and testing programs should be developed on a sound scientific basis" (p. 25). As Jaeger (1990) suggested, this would require that "procedures for establishing test standards be well documented, be based on an explicable rationale, be public, be replicable, and be capable of producing a reliable result" (p. 15). Standard 6.9 (1985) requires that when a cut score is used, "the method and rationale for setting that cut score, including any technical analyses, should be presented in a manual or report. When cut scores are based primarily on professional judgment, the qualifications of the judges also should be documented" (p. 43).

As has been suggested, the *Guidelines* may not apply in any legal sense to many educational examinations. Nevertheless, it may be helpful to realize that, like the *Standards*, the *Guidelines* require that any method used should be well documented and that a clear rationale should exist for its use. In addition, the *Guidelines* require that the standard "be set so as to be reasonable and consistent with normal expectations" (EEOC, 1978, p. 38298) and that the standard must be examined with respect to its impact, utility, and adverse impact.

Court Cases Regarding Standard Setting

In general, the courts have accepted judgments regarding the cut score (*Tyler v. Vickery*, 1975). Pyburn (1984) concluded that states may set standards where they wish because they are empowered to require high standards. Although many court cases suggest that professional judgment is acceptable, if a judge is convinced the cut scores are too high, the ruling may be unfavorable (*Richardson v. Lamar County Board of Education*, 1989). Rebell (1990), in discussing three challenges that were settled or withdrawn, pointed out the very high passing rate for these tests. As he suggested: "To the extent that fear of judicial intervention caused a lowering of otherwise valid and appropriate cut scores, increased court involvement in evaluation matters is a worrisome prospect" (p. 351).

Basically, I take the position that we should not lower the cut score of a test out of fear of a law suit.

Classifications of Approaches to Standard Setting

The most general level of classification partitioned the methods into two broad categories: state and continuum (originally proposed by Meskauskas, 1976). Another level of differentiation is between procedures that set standards and those that adjust standards (see Berk, 1986). This review, like many others, does not consider state models or models that adjust standards. Such models are not particularly useful. The models that adjust standards assume that some standard has already been identified. State models make the unrealistic assumption that the traits in question are truly dichotomous in nature. They have not been used much in setting standards. As Jaeger (1989) pointed out, "Logical use of a state model would require that a perfect score on any valid measure of competence be required of all examinees, provided the measure was totally reliable" (p. 492).

The models discussed in this review are continuum models that can be categorized as examinee-centered, test-centered, and compromise models (referred to by Berk as empirical-judgmental, judgmental, and judgmental-empirical).

It should be emphasized that this section of the paper looks narrowly at standard-setting methodologies (or models). That is, it is concerned with the tasks given the judges (whether they are judging items or classifying individuals into groups). Clearly a thorough discussion of standard-setting methodologies conceived more broadly would include *the total process* from deciding who will select the judges, actually selecting and training the judges, documenting the process, and so forth. The way the total process is implemented may, in fact, be of much greater importance than which model is chosen. These issues relating to the broad notion of standard-setting methodologies are discussed toward the end of this paper.

Continuum Models

Twenty-three of the methods reviewed by Berk (1986) were classified as continuum models. Eighteen of these were classified as test-centered (judgmental or judgmental-empirical) methods. Berk described five examinee-centered (empirical-judgmental) methods.

Of the test-centered (judgmental) methods described by Berk, most were variations or extensions of the Angoff (1971) and Ebel (1979) models. The Angoff, Ebel, and Nedelsky (1954) methods are all widely used (Cizek, 1991; Mills & Melican, 1987). (Smith & Smith, 1988, stated that the Nedelsky and Angoff procedures are the two most popular.) These three methods will be described and compared. The three most commonly mentioned, compromise models are the Beuk (1984), Hofstee (1983), and the De Gruijter (1985) methods.

This paper will contain a brief review of the examinee-centered (empirical-judgment) methods and a slightly more detailed review of the test-centered (judgmental) methods. What Berk referred to as judgment-empirical methods will be covered under the headings of "Considering Impact Data" and "Compromise Models."

Examinee-Centered Models

The two examinee-centered models that have received the most consideration in the literature are the borderline-group and contrasting-groups methods (Livingston & Zieky, 1982). A requirement of these methods is that the judges must be able to determine each test taker's knowledge and skills independent of the test score.[2] This is difficult at best, and Berk (1986) refers to this identification problem as the "Achilles heel" of these methods (p. 163). The identification of the groups has been considered essentially impossible to do for licensure examinees, and examinee-centered methods have not been used much for such tests. Some believe that teachers can make these types of identifications for students in their classes for *educational intervention*-based decisions. Thus, these methods receive greater support for classroom decisions. However, even in these cases, Jaeger (1989) pointed out that "It is not clear that teachers (or other judges) can successfully restrict their judgment of examinees to the elements of behavior assessed by the competency test . . . . Teachers are likely to be influenced by cognitive and noncognitive factors that fall outside the domain assessed by the test" (p. 497). Poggio (1984) suggested that lay people doubt the legitimacy of these methods, and Berk (1984) stated that the empirical-judgmental (examinee-centered) methods lend credence to the assumption that "teachers can already tell us who is competent" (p. 161).

---

[2] Much of the literature describing examinee-centered models suggests that individuals--with known identities--are judged independently of knowledge of either test scores or test performance (absent the scores). Livingston and Zieky (1982) would include judgments made about specific--although perhaps anonymous--examinees based on actual *test* performance (absent the score). The latter approach, because it uses information about the *items*, could alternatively be considered a test-centered method.

Test-Centered Models

All test-centered (judgmental) models require a set of individuals (judges) to make some judgments about the difficulty of the items for minimally competent examinees. Questions about the number of judges, how to select and train them, the definition of a minimally competent examinee, and other practical matters will be discussed in a later section of this paper.

Nedelsky Model

The Nedelsky (1954) approach is the oldest of the procedures and has been used considerably in the health profession, the area for which the procedure was developed. It can be used only for multiple-choice questions. Basically, the Nedelsky procedure involves asking each of a group of judges to look at each item and identify the incorrect options that a minimally competent individual would know were wrong. Then for each judge, the probability of a minimally competent student getting an item correct would be the reciprocal of the remaining number of responses (e.g., if on a five-alternative item, a judge feels a minimally competent student could eliminate two options, then the probability of such a person getting the item correct is 1/3). The expected score on the test for a minimally competent student would be the sum of the obtained reciprocals across all items. Of course, not all judges will produce the same score, so the total set of minimally competent scores for the judges are combined in some fashion. Typically the arithmetic average is obtained. Nedelsky made the assumption that the standard deviation of the judges' scores would be equal to the standard deviation of the scores of minimally competent students (that is, the standard error of the test at that point). Thus, this standard deviation could be multiplied by a constant K to regulate the percentage of minimally competent students who pass or fail. Assuming an underlying normal distribution, if we wished 50% of borderline examinees to fail, we would set $K = 0$, if we wished 84% to fail, we would set $K = 1$, if we wished 16% to fail, we would set $K = -1$, and so on. Meskauskas (1976) criticized this adjustment because he finds no support for the assumption regarding equal standard deviations of the judges' ratings and the minimally competent students.

The Nedelsky procedure is based on the assumption that a minimally competent examinee does not have or use partial information about the correctness of an individual item option. The examinee either knows it is incorrect or has no idea whether it is or is not incorrect (i.e., guesses blindly among those options not eliminated). That assumption is questionable, as we shall see later.

There have been several modifications of the Nedelsky technique reported in the literature. One modification was originally used at the University of Illinois College of Medicine (reported in Gross, 1985). In this modification, the judges were allowed a neutral "not sure" judgment, so there existed a 3-point rating scale. A further modification changed the scoring of the 3-point rating scale to avoid extreme values. Reilly, Zink, and Israelski (1984) provided another modification that placed no restriction on the probability of each wrong response. Rather, it allowed the judges to determine what each incorrect option's probability of choice should be.

Ebel Model

In Ebel's (1979) approach, the judges are asked to rate the items on the basis of relevance (four levels) and difficulty (three levels). These categories form a 4 x 3 grid. Each judge is asked to assign each item to the proper cell in the grid, and once that is done, to assign to the items in each cell a percentage correct that the minimally qualified person should be able to answer. (This percentage could be agreed to by the judges via some process, or one could proceed with each judge's values and average at the final stage.) Then the number of questions in each cell is multiplied by the percentage to obtain a minimum number of questions per cell. These numbers are added across the 12 cells to get the total number of questions the minimally qualified person should be able to answer.

Angoff and Modified Angoff Models

The Angoff (19771) method asks the judges to estimate the probability that a minimally acceptable person would get each item right. The sum of the probabilities becomes the cutoff score. As with other procedures, the judges' values are combined in some fashion, usually by computing the mean or median of the judges' values.

The Educational Testing Service (ETS) simplified this procedure somewhat by providing a seven-point scale on which percentages of minimally knowledgeable examinees who would get the items right were fixed (5, 20, 40, 75, 90, 95, Do Not Know) and asking the judges to mark this scale. This modification has not been used much. It has been objected to because it is an asymmetrical distribution of values and might bias judges' responses (Livingston & Zieky, 1982).

Another modification is to use an iterative process in which judges are given an opportunity to discuss their initial values and then to modify their choices. The iteration may also include providing additional information, such as the item p-values or the failure rate of proposed standards (see the section, "Considering Impact Data").

Comparison of Nedelsky, Ebel, and Angoff Models

A large number of manuscripts have been written comparing the more popular test-centered models. Some of these are research based, and some are based on logical analyses. Several points need to be stressed prior to reviewing this literature. First, there is no reason to expect the various methods to lead to the same standard. Second, because there is no true standard, the different standards derived from the different methods cannot be compared with respect to degrees of correctness. Third, the studies are all based within a particular context, and there is no particular reason to believe we can generalize from one context to another. For example, differences in the standards set by the different processes for high school competency tests may not generalize as to what the differences would be in educational licensure settings. Fourth, there are many variations of the three methods, and a research study using one variation may not generalize to another variation of the method.

Nevertheless, some procedure must be used to set a standard. Those who choose from among procedures (or develop new ones) should surely be informed by previous writings.

There has been substantial intermethodological research. Berk (1986) reported that between 1981 and 1986, 22 studies were conducted to compare standards resulting from the application of different methodologies. The comparisons of these methods have been described by both Berk (1986) and Jaeger (1989, 1990b). Cizek (1991) listed 23 different studies that involved comparisons of one of the test-centered (judgmental) models discussed above with some other method--usually another of the test-centered models discussed. For example, there have been 3 studies comparing the Ebel and Nedelsky procedures, 10 studies comparing the Angoff and Nedelsky procedures, 3 studies comparing all 3 procedures, 1 study comparing the Angoff with a modified Angoff procedure, several studies comparing test-centered with compromise (experience-centered) procedures, and several studies comparing test-centered with empirical procedures.

Differences in the standard set. Jaeger (1989) presented a table comparing the results of 32 contrasts from 12 different studies using different standard-setting methods. He computed the ratio of the largest test standard to the smallest, and these ranged from a ratio of 1.00 to 42.00. The median ratio was 1.46; the average was 5.30. Ratios of percentages of examinees who would fail ranged from 1.00 to 29.75. The median was 2.74 and the mean was 5.90. (It should be emphasized that these studies were not limited to comparing various test-centered models; they included compromise, examinee-centered, and empirical models also.) The most general conclusions that can be drawn regarding the research is that different methods produce different standards. Based upon my review of the research and previous summaries, it also can be tentatively concluded that the Nedelsky method typically yields lower standards than the Angoff or Ebel methods (Berk, 1986, p. 147; Jaeger, 1990a, p. 16). Most studies show the Angoff method to be between the other two with respect to the standard that is set (Jaeger, 1990a, p. 16). However, it must be stressed once more that it is difficult to generalize from these findings. For example, one study of high school proficiency examinations found that the Angoff method produced a higher standard for a reading test, while the Nedelsky method produced a (nonsignificantly) higher standard for the math test (Behuniak, Archambault, & Gable, 1982).

Differences in the psychometric properties. Several studies have compared the procedures with respect to interrater consistency (typically the variance of the judges' standards), and intrajudge consistency (typically the correlation of item judgments with empirical item statistics or the inconsistency between the judge's probability and that obtained empirically for a borderline candidate from an Item Response Theory [IRT] procedure). A general conclusion is that the Angoff procedure tends to produce the more consistent (reliable) standards (Jaeger, 1990b, p. 311). This conclusion is based on a variety of studies comparing two or more of the procedures. For example, both the Brennan and Lockwood (1980) and Behuniak et al., (1982) studies found the Angoff method produced a less variable standard than the Nedelsky method. Colton and Hecht (1981) found that rater consistency was best for the Angoff technique, second for the Nedelsky technique, and lowest for the Ebel technique. However, Skakun and Kling (1980) obtained lower interrater reliability under Nedelsky than under Ebel. Jones (1987) found that the Nedelsky method exhibited the poorest technical characteristics. His generalizability study showed that the amount of variance associated with the interaction of raters and items plus residual error was nearly 50% larger for the Nedelsky procedure than the variance for the Angoff method and over 67% larger than the variance for the Ebel method.

Again, it must be stressed that generalizations across settings are hazardous. Cross, Impara, Frary, and Jaeger (1984) found that Angoff's method resulted in a smaller variance of the judges' standards than Nedelsky's method for a mathematics test (the National Teacher Examinations), but a larger variance for an elementary education test. Following feedback of normative information, the two methods produced equal variances for the mathematics test and a smaller variance for the Nedelsky method in the elementary education test. It is important to note that research generally shows variances decrease following feedback to judges.

Although it will be addressed more in a later section, it is important to point out here that there is some disagreement about whether we should or should not strive for decreased variance across judges and whether a "large" variance is a negative indicator for how well the method is working. Logically there is no reason to believe that a method that leads to a smaller variance of judges' opinions is better (or more "accurate") than one that does not. Of course, if we want to infer a measure of central tendency from a sample to a domain of individuals, the variance tells something about the potential error of this inference. To decrease the potential error, we could (a) increase the sample size or (b) encourage the judges to become more homogeneous in their opinions or (c) do both. The problem with the second approach is that we may select judges to represent a heterogeneous population. To use a method that "forces" homogeneity of diverse opinions may not be wise.

Intrajudge consistency has generally been found to be higher for Angoff's method than for Nedelsky's method (Jaeger, 1990a, p. 17).

Psychological and logical characteristics of the methods. The three methods focus the judges' attention on different factors, invoke different cognitive processes, or both (Meskauskas, 1983). For example, the Nedelsky procedure forces the judges to pay attention to information in the response options. Neither of the other two methods require such attention. Gross (1985) has argued that if the similarity between response options were the primary factor determining the difficulty of a test item, the Nedelsky method might prove ideal. However, Smith and Smith (1988) did not find that to be the case. They concluded that "other characteristics of items hold the key to understanding item difficulty" (p. 272).

An assumption of the Nedelsky method mentioned earlier is that examinees randomly guess among the options judges believe should be attractive. As research by Melican, Mills, and Plake (1989) suggested, this is not a correct assumption.

Another limitation of the Nedelsky model is that it does not work well for some types of multiple-choice items. It is particularly difficult to use for negatively worded questions and "K-type" questions. It is also difficult to use the Nedelsky model for questions requiring mathematical computation (Livingston & Zieky, 1982). Finally, the Nedelsky method has been faulted for the restricted range of the inferred probability scale (only certain values can be obtained) (Brennan & Lockwood, 1980).

The Ebel method also suffers from some logical limitations or disadvantages. First, in a high-stakes examination, all questions should be at least in the upper two categories of importance. Further, the Ebel method involves asking the judges to rate items on two dimensions, and as Shepard (1984) stated, "judges do not seem to be able to keep the two dimensions distinct" (p. 176).

Thus, of the test-centered methods reviewed, it would appear that the Angoff method tends to provide standards between those determined by the other two methods, it is somewhat easier to use, it produces psychometrically better data, and there are not as many logical/psychological arguments against its use. We thus should expect that it would be the favored method. Indeed, the next section demonstrates that.

Preference for the Angoff Procedure

Berk (1986) stated that among the judgmental (test-centered) methods, "the Angoff method appears to offer the best balance between technical adequacy and practicability" (p. 147). Jaeger stated that "if the allied health professions are ignored," the Angoff method is "clearly the most widely used procedure" (1990b, p. 304). In another article, Jaeger (1990a) stated that:

> There appears to be a developing consensus, at least in applications outside the health professions, that Angoff's procedure produces more reasonable standards than do its

> competitors. Based on very limited results, it appears that Angoff's procedure will often produce more stable (and hence more reliable) standards than will its competitors. (p. 19)

Norcini, Shea, and Kanya (1988), all employed by the American Board of Internal Medicine, agreed that the Angoff technique is favored because of its simplicity and ease along with its reasonable psychometric characteristics. Mills and Melican (1988) reported that "The Angoff method appears to be the most widely used. The method is not difficult to explain and data collection and analysis are simpler than for other methods in this category" (p. 272).

Klein (1984) preferred the Angoff method "because it can be explained and implemented relatively easily" (p. 2). Shepard (1984), Cross et al. (1984), Rock, Davis, and Werts (1980), Colton and Hecht (1981), and Meskauskas (1986) also preferred the Angoff method. Livingston and Zieky (1982), in comparing the three methods, reported that "Angoff's method is the easiest of the three methods to explain and the fastest to use" (p. 54). Smith and Smith (1988), in their study comparing the Angoff and Nedelsky models, reported "an urge to say, 'Yes, the Angoff approach is more valid'" (p. 272). However, they did caution against too much generalizing from setting to setting.

Not all research favors the Angoff method. For example, Jones (1987), while finding that the Ebel method resulted in the highest standards and the Nedelsky method in the lowest standards, preferred the Ebel method: "The high agreement among raters, the small variance component associated with the interaction between raters and items, and the small variance estimate for rater means over the sample of raters supported the use of the Ebel method for credentialing examinations" (p. 74).

My Recommendation Regarding Test-Centered Models

The review of the literature suggests the general acceptance of the Angoff method as the preferred model, and this is my recommendation. The recommendation is based on the general reasonableness of the standard set, the ease of use, and the psychometric properties of the standard. One caveat is necessary. There is some literature that suggests the Angoff model works best when there is feedback regarding what empirical data indicate regarding the difficulty of the items and the distribution of the examination scores. Thus, it is important to consider impact data and compromise models.

Considering Impact Data

Jaeger (1978) was one of the first to suggest combining a test-centered (judgmental) model with an iterative procedure that included some empirical information. He originally had the judges make yes/no judgments about whether a person should be able to answer an item and, if not, whether that person should be denied a certificate (he was working in the setting of a high school graduation test). After the judges individually finished the task, they were given the results of the other judges and asked to review and, if they wished, to revise their standards. Finally, they were told the proportion of students who would have failed based on the recommended cutoff score and were asked to reconsider their ratings and make a final judgment. Although the yes/no judgment has not been used a great deal, there has been very wide acceptance of the iterative nature of the task providing some combination of information about other judges' ratings, item difficulty indices (p-values) for the total group or those who cluster around the cut score, and the impact of the cut score with respect to failure rate (commonly disaggregated by relevant demographic characteristics).[3]

Some argue against the use of any normative information to inform judges. They suggest that such an approach is contradictory to the purpose of criterion-referenced testing. Certainly the logic of licensure testing is that we should pass all who are competent even if that is 100% of the test takers. Leading writers now seem to agree, however, that normative data could at least be *helpful* to decision makers when used in conjunction with some other method. Jaeger (1990b) articulated well the predominant view of the profession:

> In keeping with the thesis that judges should anchor their recommendations in realistic information on the capabilities and performances of examinees who will be affected by a test standard, informing judges about the distribution of test scores likely to be achieved by typical examinees is recommended here. Judges will rightly differ in the extent to which their judgments are influenced by the distribution of examinees' typical performances, but keeping them ignorant of such typical performances is inconsistent with the objective of having test standards recommended by knowledgeable, thoughtful professionals. (p. 314)

In discussing the Nedelsky, Ebel, and Angoff methods, Livingston and Zieky (1982) argued as follows: "Because of the hypothetical nature of these judgments, we believe that these methods need a 'reality check.' If you use one of these methods, you should supplement it with some kind of information about the actual test performance of real test-takers, if you possibly can" (p. 15).

Other researchers take the same position. Typically, providing such data resulted in a lower cut score (Cross et al., 1984). However, not all studies showed such an impact (Cope, 1987; Garrido & Payne, 1987; Harker & Cope, 1988). Whether standards go up or down as a result of such information, I subscribe to Jaeger's position that we want well-informed judges and that we should not keep them in ignorance.

Compromise Models

Usually there are expectations for the number or percentage of people who will meet (or fail to meet) specified standards. A standard that everyone passed might be unreasonably low. A standard that no one passed would be equally unacceptable. The setting of an absolute standard that resulted in either extreme would not be considered either psychometrically reasonable or politically viable. This would seem to imply that standards are *implicitly* normative. Shepard (1984) and Hofstee (1973) both argued that standards are based on an often implicit understanding of what is expected--in other words, typical performance.

On the other hand, ignoring absolute standards in favor of taking some set percent of examinees who should pass or fail is equally unacceptable. This would represent a truly capricious approach to the problem in that it would avoid any reference to the meaning or content of the standards.[4] Shepard (1984) suggested that "both normative and absolute comparisons establish the boundaries within which plausible standards can be set" (p. 187).

---

[3] Given the wide acceptance of the wisdom of providing p-value data to the judges and given the general agreement that judges *should* make item judgments that are generally correlated with these p-values (maybe higher, maybe lower, but correlated), it is surprising to read that one group of critics (Linn, Koretz, Baker, & Burstein, 1991) suggested that these desired high correlations indicated that a method was not working!

There are compromise models that systematically employ both absolute and normative standards in setting the cut scores. In his discussion, Berk (1986) suggested that these methods "deserve serious attention by standard-setting researchers" (p. 153). Three compromise methods have been proposed (Beuk, 1984; De Gruijter, 1985; Hofstee, 1983). These are described briefly below. They are similar in that all require judges to set a passing score *directly* and all require that these judgments be incorporated into a model that includes knowledge of actual examinee performance. If performance cannot be estimated through knowledge of previous distributions and equating procedures, the final standard cannot be set until after the test has actually been administered and scored. However, the information from the judges can be gathered prior to the test administration (Bowers & Shindoll, 1989).

Hofstee Model

Basically, the Hofstee (1983) model identifies a range of acceptable cut scores, the outside parameters of which are compromise scores determined by both absolute and normative criteria. Using the Hofstee model, the standard-setting committee first identifies a maximum cutting score (*C* max). "This score is associated with a level of performance sufficiently high to be judged satisfactory even if, by its use, not one examinee fails" (De Gruijter, 1985, p. 264). The committee then identifies a minimum score (*C* min) "associated with a level of performance so low as to be judged unsatisfactory even if, by its use, not one examinee passes" (p. 264). Next, the maximum acceptable rate of failures (*F* max) and the minimum failure rate (*F* min) are set. The range of acceptable scores is then bounded by P1 (determined by point *C* min and *F* max) and *P2* (determined by point *C* max and *F* min). Finally, a cumulative frequency distribution, *f(c)*, of failure rates given various cut scores is determined and plotted relative to *P1* and *P2*. The actual standard is then determined to be the intersection of *f(c)* with a line defined by *P1* and *P2*.

A disadvantage of the Hofstee model is that the line may not actually intersect *f(c)*. This would occur if the judges' estimates fell entirely above or entirely below the score curve. If this is the case, the line can be extended until it does cross *f(c)* (Korte, 1987).

Beuk Model

Beuk (1984) proposed a similar procedure that specifies the line that defines the set of acceptable combinations of cutoff scores and passing percentages somewhat differently. In this method, each judge is asked to specify the minimum percentage of items that an examinee should obtain to pass and to specify the expected passing rate for that score. The mean and standard deviation of these judgments across judges are then computed. The means for the two decisions define a point that is plotted on a graph of the actual passing rate as a function of the passing score. The ratio of the standard deviation of the estimated passing rate to the estimated passing score defines the slope of a line running through the previously established point. As with the Hofstee method, the passing standard is set where this line intersects f(c). An assumption of the Beuk method is that the ratio of the standard deviations reflects the judges' degree of preference for an absolute rather than a relative standard.

---

[4] There are some situations where a purely norm-referenced approach might be the *most* useful. For example, if we have limited resources for remediation, we might set the standard for who receives remediation by determining how many individuals can be served under the budget and choose those individuals on a norm-referenced basis. North Carolina, in a system to be described later in this paper, used what was basically a norm-referenced approach for determining *who* should be placed into certain categories in the state, although the percentages were *determined* by a *contrasting-groups* procedure.

De Gruijter Model

The De Gruijter model (1985) is similar to the Beuk model. Each judge provides the preferred standard and the estimated corresponding failure rate. The judges then estimate the uncertainty with respect to the true values. These uncertainty estimates are used to form ratios that ultimately are used to define the circumference of an ellipse. The contact point between f(c) and the ellipse represents the compromise standard.

The De Gruijter method has not been used as frequently as the other two methods. It is considered more complex and not as readily understood by the judges or by those who use the judges' opinions to actually set the final cut score.

Comparison of Beuk and Hofstee Models

Both of these models are relatively easy to use. Judges understand them, and the data are easy to gather and require no complicated data analysis. Both involve judges' estimates of the test performance required for minimal competence and the passing rate of the examinees. De Gruijter (1985) suggested that the Beuk method may be easier to use than the Hofstee method. Two studies comparing the Beuk and Hofstee methods show no difference in the resulting standards (Bowers & Shindoll, 1989; Mills & Melican, 1987).

Two potential problems exist with these methods. First, as already mentioned, the standard cannot be set until the test has actually been administered and scored. This may be a political or legal problem, but it is *not* a psychometric problem. In fact, the preferred approach using any of the test-centered methods is to have judgments tempered by knowledge of the actual impact of various standards on the passing rates. If delaying the notice of the standard is not perceived as a political or legal problem, then this delay in setting the standard is not a negative.

The second potential problem is that the Beuk and Hofstee models do not require the judges to look carefully at the content of each item. Consequently, it would be preferable to combine one of these two approaches with the Angoff approach mentioned earlier. Generally we might expect the compromise models to produce slightly lower standards than the Angoff model (Bowers & Shindoll, 1989; Korte, 1987; Mills & Melican, 1987).

If the Hofstee model is used, there needs to be consideration regarding how to gather the *C* min and *C* max data. Hofstee (1983) failed to indicate precisely how *C* min and *C* max should be set. We could simply ask the judges to provide these values. Two other approaches have been tried. Mills and Melican (1987) used the Angoff method and simply used the cut score from the judge providing the highest value for *C* max and the cut score from the judge providing the lowest value for *C* min. This method seems too dependent upon the chance factor of having some judge with extreme views. Korte (1987) used the Angoff approach to ask judges to consider some minimum level of competency for *C* min and some maximum level of competency for *C* max. He concluded that we might as well use a direct estimation because it yielded similar results and was methodologically the least complex.

RECENT APPROACHES TO STANDARD SETTING

At least two factors have contributed to an increased emphasis on attempting some new (or retrying some old) approaches. One of those relates to a *limitation* of some of the historically popular methods. That factor has been the increased movement (back) to performance assessments (used here as those methods that are scored judgmentally and typically not scored dichotomously). Highly related to this is the non-unidimensionality of some assessments and the perception in some settings that conjunctive approaches are philosophically preferable to compensatory approaches to standard setting. The other impetus has been the *perceived inadequacies* (fatal flaws?) of previously popular methods. The test-centered item-judgment methods have been impugned by some who have investigated those procedures used by NAGB that resulted in standards perceived to be "too high." I will review briefly some approaches tried by the National Academy of Education, American College Testing (ACT) working as NAGB contractors, the National Board for Professional Teaching Standards (NBPTS), various states, and agencies outside of education.

National Academy Approaches

Among the concerns mentioned in the National Academy of Education Panel's report was that the standards set by the judges for NAGB were different for the extended-response items than for the dichotomously scored responses. (Whether this lack of congruence speaks to the adequacy of either method is a matter of some debate. Judges may, in fact, *hold* different standards of what individuals *should* be able to do on different formats.) This led to a concern of whether we should even use an item-by-item approach, and the NAE Panel commissioned a study to examine a "whole booklet" approach. The results of this study showed that "the minimum scale values for the basic and proficient levels obtained by the two, quite different, methods are *remarkably similar* [emphasis added]. For the advanced level, they are significantly different" (higher for the item judgments--McLaughlin, 1993, p. 10).[5] This approach did *not* result in less variability among the judges than that obtained in the NAGB approach.

In what was considered an external validity comparison of NAGB cut points, the NAE Panel gathered evidence from teachers in two contrasting-groups studies (McLaughlin et al., 1993). The results of these studies showed that lower cut scores were set by the contrasting groups than had been set by the NAGB method. If we started with the assumption that lower cut scores were better, then we might conclude that the contrasting-groups approach produced "better" results. However, Kane (1993) concluded that "the methodology employed in the study suffers from at least three major problems, each of which is a serious potential source of bias" (p. 11).

The NAE Panel also commissioned three studies in which "content experts" used an "item mapping" approach. Items were mapped onto the score scale where 80% of students in that grade could answer the item correctly. Using this procedure in mathematics resulted in "cut points for the basic level below the official NAEP achievement-level cut points and above the official cut points for the advanced level" (Shepard et al., 1993, p. 108). In reading, "with the exception of the grade 12 basic and proficient categories, content experts in reading consistently identified cut points above the official achievement-level cut points" (p. 108). The NAE "Panel did not in fact take the final results as either confirming or disconfirming of NAGB levels" (p. 109), (recall that a conclusion of the NAE Panel was that NAGB cut points were too high.) As the NAE Panel report suggests, "The approach used in the expert studies was not intended as a new or alternative method" (p. 110). Kane (1993), in commenting on the item-mapping method, suggested that it "is fatally flawed" (p. 7). The primary reason Kane made this judgment is that the choice of 80% is arbitrary. Any other percentage would be equally arbitrary. Shepard (1993b) basically makes the same point.

---

[5] The NAE Panel's report may be based on data in addition to what is found in the McLaughlin paper. The Panel's report references the McLaughlin study but states that "When the same group of judges set cut points using the two different judgment methods, whole-booklet ratings led to a substantially higher cut point for the basic level and a substantially lower cut point for advanced" (Shepard, 1993b, p. 67).

In summary, the whole-booklet method led to substantially the same cut points as NAGB ones for basic and proficient levels but lower for the advanced level. The item-mapping procedures generally led to higher cut points. The contrasting-groups approach led to lower cut points, but Kane (1993) has pointed out several methodological flaws in that study.

ACT/NAGB Approaches

The 1994 ACT/NAGB pilot studies (just completed in summer 1994) were designed for geography and history. These pilots incorporated some new features that had not been widely tried in the past. Others more intimately and extensively involved with these pilots can describe the design and preliminary results more thoroughly than I can. Nevertheless, I wish to give a flavor of some aspects of the pilots. Data have been gathered on the impact of informing panelists more completely by presenting them with performance distributions, by providing panelists with information on a whole-booklet score and by using an item-mapping exercise. Research on estimating cut scores for polychotomous items included asking panelists to provide mean score estimates for each cut point, estimating score point percentages (i.e., asking panelists to estimate the percentage of students at each cut point who would obtain scores at each point on the extended response score scale), a modification of this estimated percentage approach, and a hybrid method that combines a paper selection method and the mean score estimate method.

Results from these pilots should provide additional useful research information about standard-setting methodologies. The results will be used to determine the actual procedures to be used in the 1994 NAGB cut score determinations. One of the perceptions of several observers of the pilot studies in geography is that when the panelists used the whole-booklet method, many simply did an item-by-item analysis and tallied their results.

Reckase (1994) has recently completed a modeling study comparing a paper selection method with the contrasting-groups method. An assumption of the model he employed was that the performance of examinees was normally distributed and the performance task was scored on a six-point scale. The paper selection method (see American College Testing, 1993; Luecht, 1993) "requires judges to conceptualize students that are just at the borderline between categories. They are to select papers from a set that represents all levels of performance that students at the borderline would likely have produced" (Reckase, 1994, p. 9). The contrasting-groups method "requires that teachers first internalize the construct to be assessed and then select students that are above and below the criterion of success" (p. 9). Reckase found that the paper selection procedure provided better estimates of the standard and the percentage above the standard than did the contrasting-groups procedure. The latter procedure underestimated the standard and overestimated the number of examinees above the standard (p. 16).

Research by the National Board for Professional Teaching Standards

The National Board for Professional Teaching Standards desires to identify highly accomplished teachers through administration of a package of assessment exercises. These are "complex, distinctly-scored, multidimensional exercises" (Jaeger, 1994, p. 3). Jaeger posits that new approaches must be used because "all of the most prominent performance-standard-setting procedures have in common the expectation that the tests to which they will be applied are unidimensional, and, correspondingly, that the items that compose the tests contribute to a summative scale" (p. 3). (Nevertheless, one of the three methods researched by the NBPTS was an "extended Angoff" method.)

The NBPTS investigated three standard-setting methods identified as Judgmental Policy Capturing (JPC), Extended Angoff, and Dominant Profile. All three methods were investigated within a single study using the same set of judges and the same exercises. The investigation was of the assessments for early adolescence English language arts teachers. Only five of the nine exercises in that assessment package were used in the study. In a paper summarizing the three approaches, Plake (1994), one of the researchers, made the following comments:

> Results from the Extended Angoff method are very simple and straightforward . . . . Furthermore, the panelists report high confidence in their judgments, both for the expected scores and weights at the exercise level and at the assessment package level. (p. 7)
>
> Panelists were critical of the Extended Angoff method because they felt it was too discrete and disjointed from the overall assessment package certification decision. Furthermore, they were uncomfortable with the implicit acceptance of a compensatory decision rule. However, . . . 67% related feeling at least confident that this method would produce a suitable standard for the National Board Certification. Moreover, when asked about their level of confidence in [the] standard that resulted from their application of the Extended Angoff method, 83% rated their confidence level at least "high." (p. 8)
>
> Results from the Judgmental Policy Capturing method were substantially more complex than those resulting from the Extended Angoff method. . . . The JPC method also resulted in performance standards that were substantially higher than those found with the Extended Angoff Method. (p. 7)
>
> Panelists were likewise critical of the Judgmental Policy Capturing method, again feeling that the method was focused more on the component parts. . . . When asked how confident they felt that the JPC method would produce a suitable NBPTS standard, only 42% indicated they were confident or very confident. . . . However, . . . 84% reported having at least high confidence in the standard that resulted from their JPC ratings. (p. 8)
>
> When asked to identify which of these two methods, Extended Angoff or JPC, gave them the greatest confidence in producing a suitable standard, . . . 8% preferred the Judgmental Policy Capturing Approach, 33% showed preference for the Extended Angoff method. (p. 9)

> Although no formal evaluative information was obtained from the panelists directly about their reactions to the Dominant Profile Method it seems reasonable to conclude that they would be more confident and comfortable with the policies resultant from this approach. (p. 9)
>
> The Dominant Profile method, as currently designed, has a substantial order confound. It is not clear how the method would work without the preliminary orientation provided by the Extended Angoff and JPC methods. (p. 12)
>
> From the results of this study, it is clear that the JPC and Dominant Profile methods, perhaps in tandem, show promise for providing a useful method for setting performance standards on complex performance tasks. (p. 12)

The researchers who conducted the single study investigating these three methods are deservedly recognized as some of the most capable in the field. Their findings are somewhat positive regarding all three methods, and further research should be encouraged. Nevertheless, as they would admit, the research is not conclusive. Two major limitations need to be mentioned. First, only five of nine exercises were used in their research. Certainly we might expect the JPC and Dominant Profile methods to become exponentially more difficult as more exercises are added. Second, there was clearly a confounding of order with method in their research. There is no good way to infer the impact of this order and whether, for example, the Dominant Profile method would have worked as well without doing *both* the JPC and Extended Angoff methods first. (Putnam, Pance, and Jaeger, 1994, advocate using the JPC method as an enabling procedure for the Dominant Profile method.) Another weakness of the study is that there was not a formal evaluation of the Dominant Profile method.

Finally, it should be pointed out that their research did not investigate what differences might exist if we aggregated scores on single dimensions across exercises and then judged the dimension scores versus aggregating all dimension scores *within* an exercise and then judging the exercise scores. Should we expect the same standards to be set under the two conditions (in terms of percentage and commonality of examinees passing)? If not, which method is to be preferred? If, as we typically find, method variance predominates over dimension variance, what does this say about either the validity of the assessment or the standard-setting methodology that should be employed?

States' Approaches

Kansas

Poggio and Glasnapp (1994) have reported on a new test-centered (judgmental) method for setting multilevel performance standards on objective or constructed response tests in Kansas. They suggest that this approach:

> (1) overcomes many of the practical and psychometric problems associated with the Angoff and Ebel methods, (2) can be used to set multiple cut points on a score scale, (3) may be readily and efficiently implemented with assessments that use objective or constructed response items or both, and (4) allows participation in the standard setting process of persons who may not be educators or not necessarily familiar with the instruction of individuals with whom the examination will be used. (p. 1)

This new method requires each judge "to specify the minimum acceptable score performance *distribution* . . . s/he would define as just barely acceptable in order to identify the referent group of examinees as having demonstrated performance in the target classification" (Poggio & Glasnapp, 1994, p. 4). The process is a two-step activity where participants first review each test question and provide a judgment regarding each item's cognitive demand or complexity. This step is to ensure that judges are familiar with the content, substance, and format of the assessment. Next the judges are asked to specify the score distribution of a group of 100 examinees who are "just barely meriting the classification label to be assigned" (p. 6). The cut score for each judge is the average of that judge's distribution, and the overall cut score is the average of the group of judges.

The researchers report that the new procedure had an average interrater reliability of 0.89 (compared to 0.80 for the Angoff procedure); the variances of the judges cut scores and the cut scores themselves were lower in this method than in the Angoff model); and the cut scores from multipoint performance settings were comparable to the cut scores for dichotomously scored items. Although the results of this single study are favorable, the task asked of the judges seems quite complex. Further, it is not clear to me why this procedure can be used by noneducators any better than what seem to be less demanding procedures. There should certainly be more research on this method before it is widely adapted.

Maine and New Hampshire

Kahl, Crockett, and DePascale (1994) report on two methods they have used. A Student-Based Constructed Response (SBCR) method was used in both New Hampshire and Maine, and an Item-Based Constructed Response (IBCR) method was used in Maine. The SBCR approach places students on a Rasch ability scale, and the IBCR approach places score points for individual items on a Rasch scale. In the SBCR approach, judges look at sets of student responses' that have been placed in folders based on quarter logit estimates of the students' Rasch ability estimates. Judges are asked to assign the students' responses to proficiency levels. The authors suggest the SBCR approach is like the holistic scoring of student portfolios in which many samples of student work illustrate the students' capacities. (They seem to think of this as a positive even though scoring on student portfolios is remarkably unreliable.) Judges in Maine were more confident in using the SBCR approach. The 1994 paper does not present actual results, and the authors note that "many additional analyses of the data will . . . shed additional light on the impact on cut scores of such factors as the method used, the background of judges, and the extent of exposure of the judges to the test questions" (p. 11). Nevertheless, the authors conclude that "certainly the methods of standard setting described herein are more appropriate than traditional methods considering the current status of multiple-choice testing" (p. 10). I was unable to discern the bases for this confident statement.

Connecticut

Smith and Greenberg (1990) report on a strategy used in Connecticut for setting standards on a performance-based observational test for beginning teachers. There are several aspects of this assessment that are different from the usual multiple-choice test. The assessment is based on observers giving an acceptable or unacceptable rating to the behaviors associated with each of 10 indicators. Thus, one standard of performance is made at the observational level where a judgment is made regarding whether the teacher has demonstrated behaviors to earn an acceptable rating. Also, the 10 indicators are *not* considered samples of some larger domain. They constitute the set of critical dimensions of teaching.

The actual assessment data consist of three observations in the fall and three in the spring for each of the 10 indicators. Thus, "the goal of standard setting is to come up with a formula for combining the 60 separate *Accept/Unaccept* ratings into a single pass/fail decision" (Smith & Greenberg, 1990, p. 2). We could use either an *observation-focused* approach or an *indicator-focused* approach. (This is somewhat analogous to using either an exercise approach or a dimension approach for the NBPTS assessments.) It was determined that the *indicator-focused* approach was more consistent with the purposes and goals of the teacher observation instrument because the indicators represented the essence of teacher competence.

We could use a completely compensatory, partially compensatory, or noncompensatory formula. In addition, two standards had to be established: acceptable for a provisional certification or, if not, acceptable for a one-year extension.

The original design called for a panel of subject-matter experts to respond to nine different questions designed to answer questions about which model to use and what standards should be expected. The results were to be tabulated and reported to the panel along with some information from a pilot study and a job analysis. The plan was *not* to achieve consensus, but to allow panelists to reconsider their judgments.

Although the original plan called for the panel of experts to choose whether to use a compensatory or noncompensatory model, this decision was instead made at the policy level. The policy decision was that performance on *each* indicator was critical for provisional certification. Therefore, all aspects of the design that were related to the compensatory model were eliminated. As a result of some mini-simulations, the design was simplified further, and eventually only four questions were asked:

1. For each of the 10 indicators, circle the number of *acceptable* ratings a beginning teacher must obtain in order to be judged competent on that indicator.

2. Circle the minimum number of indicators on which a beginning teacher must be rated as competent.

3. (and 4). Circle the percentage of all beginning teachers you believe will meet the standard for provisional certification (or for an extension).

This was an interesting paper describing a detailed design. Although I agree with the researchers that the task demands of the original design were high, I think it was somewhat unfortunate that much of the design was not implemented due to policy decisions. Although I support the position that all standard-setting decisions are essentially policy decisions, I think it would be useful to gather information from a panel of experts regarding the compensatory vs. conjunctive models, just as it is useful to gather information from the panel regarding suggested standards.

North Carolina

North Carolina wished to set standards for four levels of performance based on a grade level concept: Level I students were not at grade level, Level II students were marginal performers at grade level, Level III students were solid achievers, and Level IV students were above grade level. North Carolina personnel first performed what may be considered a contrasting-groups approach by asking all teachers in grades three through eight to categorize their students in these four levels. They found that adjacent contrasting groups overlapped greatly in achievement on the state tests. Rather than using the intersections of the frequency distributions to set cut scores, North Carolina simply used the proportion of students in the state classified in each category to determine the cut scores. A replication study the next year showed these proportions to be very stable.

I would consider the approach used in North Carolina to be a combination of a contrasting-groups approach and a norm-referenced approach. The intersections did not determine the cut score, but neither was the approach pure norm referenced. That is, the teachers did not directly state what proportion they believed should be categorized in each group; they actually categorized all students, and the resultant percentages determined the standards.

Noneducational Approaches

Although different decisions are reached in noneducational settings than in educational settings, the standard-setting literature has not always differentiated procedures for testing purposes. However, there are some differences in procedures. For example, who is on a standard-setting committee may differ. For a licensure or employment decision, one would not think of asking the general public to be judges although this has been done with respect to educational standard setting. Another possible difference may be the relative costs of false positives and false negatives. But neither of the two differences mentioned would necessarily lead to different standard-setting models. Perhaps one of the major differences is that there has been somewhat more experience with performance assessments in noneducational settings. The few examples I wish to discuss involve setting standards on nonmultiple-choice tests.

Klein (1991) has described an approach to setting standards on constructed-response licensure tests, which he suggested may be appropriate for various licensure tests such as for teachers or lawyers. The procedure has been used on a bar exam. Basically, the procedure works by comparing the regular readers' scores to evaluations by an independent panel of experts on free response tasks. In the bar exam Klein describes, each of 12 essay questions is graded in terms of its performance. The sum of the 12 raw scores is a candidate's raw essay score, which is converted to a standard score. After all essays have been graded by the regular readers, answers are selected to each question where scores range from high to low. Five expert panel members independently classify each answer into one of four categories (clear pass, marginal pass, marginal fail, or clear fail). The panelists must discuss their ratings and reach consensus on the categorization of each answer. These classifications are used to determine what score assigned by the regular readers best differentiates between the marginal pass and marginal fail categories. Based on this score, the percent of candidates who pass the question is determined. This is done for each of the 12 questions, and the median of the 12 passing *rates* is designated as the percent passing the exam. The total score that produced this median passing rate is the passing score for the exam (which is a combination of essay and multiple-choice questions).

Klein (1991) pointed out several key features of this process. One is that the panelists do not know whether their classifications will yield a high or low passing rate. Another feature is that the standards are set on the essay portion of the exam (the portion that lawyers believe to be the most valid). Klein has monitored this process by asking the regular readers what score they considered passing on the questions they read. Aggregating these scores in the same way as described above results in standards that were "always within 10 percentage points of the actual passing score" (p. 5).

Norcini et al. (1993) have described standard-setting approaches that have been used in assessments that employed multiple station assessments of clinical skills. In clinical station assessments, typically continuous rather than dichotomous scores are obtained for each item, and only a few items are given at a single station. In the Norcini et al. research, the stations were scored dichotomously by having experts go through the checklists obtained from the station assessments and sort them into two groups based on the candidates' performances. (Recognize that this is not the same as the typical contrasting-groups approach where the groups are formed independently of the assessment data.) Then the mean of the distribution of scores of the assessments assigned as failures was used as the cutting score for each station. The Angoff approach was used with these dichotomous scores. These Angoff values were multiplied by the mean of the distribution of candidates scores given a "pass," and thus an overall standard was determined.

In a follow-up to the Norcini et al. research, Rothman, Poldre, Cohen, and Ross compared the results of scoring and standard setting with the dichotomous and continuous scales. In their study, there were 29, ten-minute stations (all contained patient encounters and 21 linked the encounters with postencounter written exercises). The 29 stations' assessments were each scored in a continuous fashion, with scores ranging from 1 to 20. Also, physicians observed the encounters and completed exercises. Contrasting pass/fail groups were obtained for each station by having the examiners record whether each of the candidates had or had not performed adequately (again, this is not the same as the typical contrasting-groups approach). For each station, a cut score was defined as the point of intersection between the score distributions of the passes and failures. The mean of these 29 scores determined the test standard on the continuous scale. This standard was compared with one obtained using the dichotomized station cut scores (where the total test score was the number of stations passed). A borderline group of 10 candidates was obtained based on their continuous scores (5 immediately above and 5 immediately below the continuous score standard). Angoff-type values were obtained for the stations using these assessments to help define minimally competent.

The passing rates using the two methods were 81.4% for the continuous method and 78.6% for the dichotomous. Of the 16 candidates who failed under one or the other approach, 12 were failed by both. The authors concluded that the psychometrics did not favor either alternative because both sets were satisfactory. However, the consequences of the choice taken were significant because of the extent of disagreement between the two scales in the classification of failures (Rothman et al., p. 7).

Julian (1993) wrote a paper designed to "stimulate discussion of the assignment given to the standard-setting judges, and of how they can account for the difficulty of the task while fulfilling that assignment" (p. 1). She proposed a two-stage content-based procedure for setting standards on performance examinations. In her terminology, a task may be divided into items, which are frequently the unit of scoring within a task. However, these items are often accumulated only within a task, and the task scores are used to build an examination score. Julian suggested that setting task-level standards at the item level through procedures such as Angoff's "fails to capitalize on an advantage of performance examinations, which is the ability of the judges to grasp the task as a whole, and to make a holistic judgement about performances" (pp. 3-4). She proposed that judges look at representative performance and make pass/fail judgments about each task. However, this gives the judgments about minimum competence on the task (MCT), whereas we might wish to define a score that represents minimum competence on the *domain* (MCD). If the average difficulty of the tasks in the examination is equal to that of the tasks in the domain, the distinction is unnecessary. However, if they are not, the two standards will differ. Julian argues (correctly in my opinion) that the standard-setting process for performance examinations often does not make the subtle and important distinction between the performance expected of an examinee who is minimally competent on the TASK (MCT) and that expected of an examinee who is minimally competent in the DOMAIN (MCD) (pp. 6-7).

As Julian (1993) suggested, one way to tell which standard the judges have set is to compare the percentage of examinees failing different tasks. If the "standards are set at the MCD, then task difficulty should not impact an examinee's probability of passing that task" (p. 7). Julian thus proposes a two-step method. First, the judges would set the standards to represent competence on the task. Second, the "judges would identify the probability of the examinee who is minimally competent in the domain 'doing ok' on the task" (p. 8). Using Item Response Theory (IRT), this probability provides an ability level. Once standards have been set for the tasks, a standard can be set for the total examination either by dichotomizing the task scores and requiring a minimal number of passes or by using continuous task scores and requiring a minimal total score. Obviously the first approach does not allow any compensation across tasks, and the second approach assumes all tasks are scored on the same metric.

Summary of the Literature Review

As the review of the literature indicates, a multitude of methods exist for setting standards. Those that are considered most frequently are continuum models. Within this category, those used most are test-centered models. Historically, the preferred test-centered model has been the Angoff approach because it is easy to understand and implement, provides reasonable standards, and has the best psychometric characteristics. There is some debate currently about whether we should modify this for nondichotomous items or use another approach.

Logic and research show that judgments need to be tempered by reality. One very common way to incorporate empirical data is to use a modified Angoff approach that provides the judges with data regarding the item-difficulty values or the impact data (failure rate) or both. This is typically done in an iterative fashion by obtaining the Angoff judgments on the first iteration and then providing information regarding item/test data on subsequent iterations.

Another set of approaches is called compromise models. Of these, both the Beuk and Hofstee models are popular and provide similar results. The choice between them is basically a matter of individual taste, although the Beuk method makes a questionable assumption regarding the meaning of the judges' standard deviations, and the Hofstee model may require an extrapolation of a line in order to set a standard. If the Hofstee approach is used, I would recommend a direct estimation of *C* min and *C* max.

A recent report by the NAE Panel (Shepard et al., 1993) has called into question the popular Angoff and all other item-by-item approaches, although the Panel's discontent is not shared by many. Their "evidence" suggesting that the Angoff method is flawed and that other approaches are to be preferred has been questioned. Although the NAE Panel opined that NAGB standard was "too high," the NAE Panel-commissioned research has indicated that a whole-booklet approach set similar standards (although lower at the advanced level) and that an item-mapping method produced even higher standards. (The NAE Panel research using the contrasting-groups approach led to lower standards, but the methodology was criticized strongly by Kane, 1993.) Recent writings (e.g., Kane, 1994; Reckase, 1994) suggested that there are bias problems in the contrasting-groups approach. The NAE-commissioned research did indicate that setting three cut scores as was done by NAGB for the National Assessment of Educational Progress scores (as contrasted with the more usual process of setting one standard) did *not* seem to be a problem.

There is some evidence from NAGB approaches that the standards set by the Angoff approach on dichotomously scored items are not at the same points as the standards set on nondichotomously constructed responses. There is debate about whether this suggests that at least one of the cut scores must be "wrong," because, after all, NAGB process was to ask how students should perform, and the judges might really have different values about "should" performances on these two item types.

There is general agreement among experts that more is known (or at least that more research has been conducted) about how to set scores on dichotomous than polychotomous items. There has been some recent research on standard setting for polychotomous items to redress that lack of balance. In general, this research has still employed an item-based approach *if* we conceptualize the task as an item. However, there seems to be a preference for not looking at the subscorable parts of a task as items and setting standards on those "items." The task judgments have typically included looking at a sample of responses to the task. Judges may sort these task responses into contrasting groups (recognize that this is different from a contrasting-groups approach where the groups are determined by criteria independent of the assessment), estimate the mean score for borderline candidates, estimate percentages of borderline candidates who will obtain each score on the extended task scale (this would probably work only for tasks with reasonably few scale points), or some combination of these procedures.

A number of approaches have been used in aggregating judgments across tasks into a total standard. We might set standards on tasks and then convert task scores to pass/fail scores and aggregate by setting a minimum number of tasks that need to be passed. Conversely, we might use the task cut scores on the continuous scale when aggregating to a total standard. Some research has supported the continued investigation of combining a Judgmental Policy Capturing method with a Dominant Profile method. One study, although requiring judges to examine and make judgments about the individual item difficulties, actually reported success in having judges estimate a distribution of total scores for a set of 100 borderline candidates.

Researchers have used different approaches to setting a total standard when a total test is composed of different item formats (some of which lead to dichotomous scores and some to polychotomous scores). Klein (1991), for example, set the cut score on the performance scale to obtain a passing *rate* and used this same passing rate for the combined total test. NAGB approach has been to scale all the items onto a single scale and set the standard on that scale. Obviously we could set the total standard on only the dichotomous items but that would probably not be politically acceptable. Of course, we need not scale the two types of formats together. We could scale them separately, set separate standards, and require that one pass both types of formats (e.g., a multiple-choice exam measuring mechanics of writing and an essay exam).

AGREEMENTS ABOUT THE METHODOLOGY OF STANDARD SETTING

Some points are listed below that I believe are generally agreed to regarding standard setting. Most of these points are backed by considerable literature. Others simply seem to be a matter of logical necessity.

1. Standards are judgmental; there is no right answer as to where a standard should be set (although some believe there are wrong answers).

2. Context matters. A method that may be best in one context may not work well in another context.

3. Test-centered methods, if used, should be supplemented by information on item statistics and impact data.

4. Impact information should be based on real test data rather than on field test data.

5. If an item-by-item approach is used, the Angoff method is the preferred test-centered method for dichotomous items.

6. Test-centered methods cannot be used without judges at least seeing the items that comprise the assessment. The more thoroughly they understand the items, the better they should be able to do their job.

7. An item-by-item judgmental approach *forces* greater scrutiny of the items than, say, would necessarily occur under a whole-booklet approach.

8. Examinee-centered methods require judges to determine each test taker's knowledge and skills independent of the test score.

9. For some purposes (e.g., licensure decisions) there is no group of judges that can fulfill the requirement mentioned above. For other purposes (e.g., classroom intervention decisions), there is debate about whether judges can perform this task.

10. The contrasting-groups method is the most popular of the examinee-centered methods.

11. Whether judges use an examinee-centered or a test-centered approach, they cannot set standards without some understanding of what it means to be at the borderline for a specific categorization. (There is room for disagreement about how general or specific this understanding should be.)

12. The understanding of borderline should not be changed after the standards are set.

13. Judges must understand the method that is being used but not necessarily all the math that has gone into obtaining scaled scores, etc.

14. Setting standards for polychotomous items (tasks) is not as well researched as is setting standards for dichotomous items.

15. Judges should see a sample of examinee papers (preferably real test papers but at least field test papers), or some record of their responses, prior to setting standards on performance assessments.

16. Based on previous statements, it follows that standards should not be based on *only* a consideration of scoring rubrics. (If the rubrics were developed based on examinee responses and a careful consideration of what examinees should know and be able to do, the rubrics may provide some basis for the standards.)

17. Subjective scoring of performance is corruptible. Once standards are set, the rigor of subsequent scoring may change because scorers (readers) have a conscious or unconscious preference for a different passing rate. (There is at least some anecdotal evidence that this has actually happened in some states where principals were involved in scoring observations for purposes of making certification decisions.)

18. Setting a standard by totaling task scores is inappropriate if a policy requires conjunctive standards.

19. Judges cannot really understand the passing rate impact of conjunctive standards (either across tasks within a single assessment or across assessments--e.g., having to pass all four subject-matter tests to receive a high school diploma). To assist the judges, it is essential that the judges have impact information on the conjunctively set standard. They also should be given information regarding how that impact would change if different, separate (prior to combining in the conjunctive model) task (test) standards were set or if the conjunctive rule were changed or both (e.g., 12/12 vs. 11/12 vs. 10/12).

DISAGREEMENTS (OR CONFUSION) ABOUT STANDARD SETTING

The following points are stated as propositions. I believe there is often strong disagreement among scholars about the truthfulness of these propositions. Readers should *not* infer that I believe that the propositions are true. My current, somewhat tentative, opinions can best be inferred by reading the parenthetical statements.

1. Empirical findings regarding what examinees *can* do, independent of the assessment, is evidence regarding the correctness of a standard based on what judges *believe* examinees *should* be able to do. (This depends on whether "correctness" refers to the correct reporting of the belief or to a correct belief. From a strictly logical basis, correctness of *reporting* a *belief* about what *should* be is independent of what *is*. However, the belief may be unwise [incorrect]. I believe that judges should be given "can do" information to help them formulate their beliefs about "shoulds.")

2. Variance across judges' ratings is a relevant criteria in judging the adequacy of the method. (Obviously it is relevant to the amount of error that might be made in inferring from the standard set by a group of judges to the standard that would have been set by the total population of judges from whom they were drawn. Although the standard error can be made smaller by increasing the number of judges, it is not necessary to decrease the variance of their opinions. There is no reason to believe that a set of judges--especially if intentionally chosen to represent a heterogeneous population--would come to very homogeneous views.)

3. Item-by-item approaches are fatally flawed. (Twenty years of research does not support this position, and I do not believe that most scholars agree with it, although it is apparent that the NAE Panel still entertains this as a strong possibility.)

4. It is reasonable to have nonexperts make judgments about standards. (This should not even be considered for purposes such as licensure decisions. I do not see how the

general public could make decisions about what fourth or eighth graders should know before being promoted--or even classified as advanced, proficient, or basic. I see *some* logic in the general public being represented on a panel setting a standard on what a high school graduate should know or be able to do [either to graduate or simply to be classified]. However, from a purely methodological point of view [ignoring politics], I would always prefer the panel to be experts on the domain being assessed and, if children are involved, on the developmental level of the children being assessed.)

5. If we cannot set standards in a scientific manner, we should not set them at all. (I generally disagree with this. First, as this conference will no doubt demonstrate, there can be disagreement about whether a procedure is scientifically acceptable. More important, we make dichotomous decisions. If we do not set standards on assessments to assist in this decision making, we likely will make poorer decisions. Of course, we do not need to categorize for all decisions. Sometimes reporting the score on a continuous scale is preferable.)

CONSIDERATIONS OTHER THAN CHOOSING A MODEL

In addition to choosing a particular standard-setting model (or set of models), a number of other specific decisions remain to be made. These include how the judges should be chosen, how many should be involved, how they should be trained, whether they should meet as a group, how their tasks should be monitored, how separate judge's decisions should be aggregated, what should be documented, and what information should be taken to the policy board that makes the final decision. This section of the paper addresses some of these details, but it is not intended to obviate the need for a knowledgeable individual (or committee) to plan and execute the actual process.

Choosing Judges

Basically the decisions regarding choosing the judges involve how and how many. Jaeger (1990a, 1990b, 1991) has spoken to this issue as thoroughly as any writer. Obviously, the judges must be qualified and credible individuals. For licensure exams, the judges would ideally come from a set of already-licensed individuals. If the exam is for a new teacher's license, the judges should be chosen from among a set who have expert knowledge in the content as well as a good understanding of the tasks a teacher will need to perform. For examinations given to public school students, I would prefer the judges be educators with knowledge and experience both in the subject matter being assessed and at the grade level of the students being assessed. (I would not rule out the general public for what high school graduates *should* know.)

Once a population of judges has been identified, a representative sample should be chosen from that population. The sample should be representative of various demographic factors such as geographic region, ethnicity, and gender. As Jaeger (1990b) pointed out, a "difficult practical problem in selecting a representative sample is assembling a list of all members of the desired population of judges" (p. 300). Certainly in many cases, we simply cannot assemble a list of all qualified judges from which to sample. The best that can be done is to make sure that no judge is selected who is not qualified and to strive for as much representation across demographic variables as possible. However judges are chosen, the *Standards* state that "the qualifications of the judges should be documented" (AERA/APA/NCME, 1985, p. 43).

The number of judges involved in a standard-setting method is important because the standard error of the mean (or median) is based on the sample size. The standard error of the mean is a measure of the standard deviation of the sampling distribution, or how much the standard is expected to fluctuate from sample to sample. This should be small if we desired to claim that the standard set represents the judgment of a population of judges rather than just the sample that happened to be chosen. Jaeger (1991) suggested that a reasonable goal is to have a standard error of the mean that is no larger than one fourth the standard error of measurement of the test. Making various assumptions about both the standard error of the test and the standard deviation of the judges, he showed that the number of judges needed to meet that goal might reasonably be between 13 and 21. In general, 20 representative judges seems a reasonable number.

Training Judges

Typically, judges have never been involved in a formal standard-setting task, and it should not be assumed that they will know how to proceed. They must be thoroughly trained to do the job. Basically, they need to understand the general task, which includes the purpose of the standard-setting process; they must have a shared understanding of what it means to be minimally competent; and they must understand how to do the particular standard-setting method that is to be employed. It is generally preferred to have the judges trained at a group meeting (Cizek, 1991, p. 6), although they should receive materials in advance of the meeting so that they may obtain some background information regarding the standard-setting process.

At the group training meeting, the first step should be to make sure the judges understand the purpose of the examination and the need for setting a standard. If, for example, the purpose is to set a standard for a licensure examination, the judges should understand that the purpose of a licensure examination is to protect the public. It would probably be useful to discuss the notions of false positives and false negatives and to share views regarding their relative costs.

In setting the general context of the meeting, it is also useful to discuss how the content specifications and test items were developed. As Mills et al. (1991) pointed out:

> In this way, concerns of the judges about the content of the test can be addressed and separated from the rating of the actual items. . . . Judges should understand that their task does not include establishing the need for the test, selecting content to be covered, or writing/reviewing the actual test items. Rather, the judges' task is limited to assessing the difficulty of test items for the minimally competent entry-level individual. (p. 8)

There needs to be a discussion regarding what it means to be borderline. For licensure decisions, Mills et al. (1991) suggested one method that "begins with a global definition of the skills routinely required for practice and to develop refinements of that definition to address (a) typical entry levels and (b) minimally acceptable levels for the beginning practitioner" (p. 8). Arrasmith and Hambleton (1988) have suggested two general strategies for obtaining an understanding of minimally competent in a profession. Both strategies involve having the judges review the domain specifications and objectives and having a discussion about the necessary skills for an entering professional. All these general suggestions seem generalizable to settings other than licensure.

After sufficient discussion regarding the purpose of the examination and an understanding of a borderline individual, there should be specific instructions regarding how to complete the particular tasks. If, for example, an Angoff procedure is employed, judges should understand whether they are to estimate the proportion of the minimally competent who will *know* the answer or the proportion who will *get it correct*. The second proportion will be different because of a guessing factor. If a correction for the guessing formula is *not* used in scoring the test, the judges should estimate the proportion who will get the answer correct. It is advisable to have some practice exercises that the judges should complete and then discuss (Reid, 1991). A discussion of the answers should include the factors that make items more or less difficult (e.g., negatively worded stems, *K*-item formats, and attractiveness of the distractors). If item-analysis data are available on these practice items (e.g., from a field test), this information should be shared with the judges.

It is generally considered advisable for the judges to actually take the test under regular test conditions prior to making their judgments. The correct answers should then be given to the judges so they can discern how they did. This activity forces them to carefully consider each item, and it gives them a better understanding of the conditions under which test takers operate.

The judges have to be trained to understand any data that are provided to them, and they need to understand that such data do not take the place of their judgments, but rather assist them in making judgments. Special care needs to be taken to explain the effect of using a conjunctive model on the overall passing rate.

Iterations and Monitoring the Judges

As mentioned earlier, most experts on standard setting using a test-centered approach prefer an iterative procedure. After training, the judges individually rate the items. These ratings are aggregated across items and the proposed standard for each judge is determined. These will differ by judges, and the distribution of standards should be shared and discussed. Subsequent iterations would include more data (if available) about the difficulty of the items and the impact of various cut scores.

A typical finding is that if judges are allowed to reconsider their initial standards, the judges' standards become more homogeneous (Cross et al. 1984). Busch and Jaeger (1990) provided the following statement with respect to allowing judges to discuss and reconsider their initial recommendations:

> In summary, then, allowing judges to discuss their initial standard-setting recommendations enjoys some support in the social psychological literature and also can be supported through the logical claim that discussion results in judges becoming better informed about relevant bases for judgment. In addition, a limited empirical literature suggests that discussion reduces the variability of recommended test standards, with concomitant improvement in the reliability of recommended standards. (p. 149)

Although favoring discussion and reconsideration, the profession does *not* advocate forced consensus (Brennan & Lockwood, 1980; Jaeger, 1988).

Judges should be monitored with respect to both interjudge and intrajudge variability (occasionally referred to as consensus and consistency). As mentioned, interjudge variability can typically be reduced by the iterative process and informing judges about how the other judges are rating. Intrajudge variability can be monitored through correlating the judges' ratings with the actual p-values of the items. Judges should be informed, for example, if they rate the probability of a very easy item much lower than the probability of a very hard item. In addition to correlational analyses, Generalizability Theory, IRT, and the Modified Caution Index procedures have been used to assist in monitoring intrajudge consistency (Friedman & Ho, 1990; Jaeger, 1988; van der Linden, 1982). It is comforting to know that research indicates that techniques designed to improve consensus also improved consistency (Friedman & Ho, 1990).

Aggregation of Judges' Ratings

Regardless of whether iterations have been used, it is common not to force total consensus among the judges, and so a decision remains regarding how to combine their responses into a single recommended standard. Typically, the mean of the individual judges' standards has been used, but we could also use a median, a trimmed mean, or any one of the three procedures with a subset of judges who have had the greatest intrajudge consistency. We could also weight the judges' individual standards based on intrajudge consistency.

Research on eliminating some judges' responses or weighing judges' responses differentially suggests that such a process will improve the reliability of the cut score (Maurer et al., 1991). However, it may not change the actual standard (Friedman & Ho, 1990). Friedman and Ho (1990) eliminated judges on both the consensus (outlier) method and the consistency method. They found that the two methods eliminated somewhat different judges (there was some overlap) but that "the cut score for the groups selected by the two methods did not change" (p. 11). Jaeger (1988) eliminated judges on the basis of a modified caution index. He found that such screening can affect the standard but that it did not result in reducing the standard error of the mean. He opined that:

> Achieving consensus on an appropriate standard for a test is an admirable goal . . . , but it should not be pursued at the expense of fairly representing the population of judges whose recommendations are pertinent to the task of establishing a workable and equitable test standard. To eliminate the recommendations of some judges *only* because they differ from those of the majority is antithetical to the more fundamental goal of seeking the informed and reasoned judgments of one or more samples of judges who represent the population or populations of persons who have a legitimate stake in the outcome of the testing program under study. This argument suggests that procedures for screening judges be used cautiously and thoughtfully, if at all. (p. 29)

If we are concerned that some judges may give an "incorrect" standard because they simply have not understood the standard-setting task or were incapable of performing it (and *not* because they had a different opinion of what the standard should be), this could be determined in the training session. Individuals could be trained on sample items and not allowed to go into the actual standard-setting session until they had demonstrated sufficient intrajudge consistency.

In my opinion, no judge's standard should be eliminated from the set of data to be aggregated. It may seem contradictory, but it does not necessarily follow that we cannot aggregate that data set through use of a median or trimmed mean. The mean has the advantage of having the smallest standard error. It also allows all judges to believe that their opinions count. It has the potential disadvantage, however, of being affected by extreme scores. Thus, we might argue for a median or trimmed mean. Livingston and Zieky (1982) suggested that:

> If you are going to use the trimmed mean for averaging the scores, you should let the judges know this fact before you calculate the passing score from their judgments. Otherwise, the judges with the highest and lowest standards may suspect that you are discriminating against them. (p. 23)

I would make the same point about the median (which is a very trimmed mean).

My view is that the final decision regarding the standard is made by the board that has the authority to make such a decision. It is not necessary to choose only one way to aggregate the data from the standard-setting committee (which brings forth recommendations rather than making a final decision). I believe both the mean and median should be computed. This information, along with the individual judges' standards and the standard deviation of those standards should be taken to the board.

Documentation

All processes engaged in and the decisions reached in any standard setting for a high-stakes test must be thoroughly documented. The following list provides a *sample* of procedures and decisions that should be documented:

1. What standard-setting procedure(s) is (are) to be used? Who made the decision, what was their authority, and why did they choose the particular procedure(s)?

2. How were the judges selected? Were they drawn randomly (or stratified randomly) from a defined population? What was the acceptance rate among the judges selected? What were the qualifications of the judges? How were the judges distributed with respect to various demographic factors? What was the basis for deciding on the number of judges?

3. What was the training process? Is there documentation regarding how confident the judges felt about their task?

4. What were the initial and any subsequent iterations of the judges' recommendations by item and by final cut score?

5. What was the initial and subsequent interjudge consensus and intrajudge consistency?

It should be understood that the above list is suggestive only; it is not meant to be the total list of processes/decisions that should be documented.

Summary of Additional Considerations

Setting a standard involves many considerations in addition to the particular standard-setting method to be used. If judges are to be used, it must be determined how to identify the appropriate potential set of judges, how to choose from that population, and how many judges need to be chosen. Judges should be demographically representative and of sufficient number (say 20) so that we can feel confident regarding the ability to generalize their recommended standard to the population of judges from which they were chosen.

The judges must be thoroughly trained in the method(s) they are to use. This involves understanding their task and practicing on sample items. The judges should use an iterative process, with additional information on how the other judges rated the items and whatever empirical evidence on the items is available added on subsequent iterations. The intrajudge consistency of each judge should be monitored. Whatever judges' ratings are aggregated to produce a standard (or set of standards) to take to a policy board, this information should be available to the judges at the outset. All procedures should be documented, and the documentation should be maintained in a file at least until a new standard is set for the exam.

PREPARING POLICYMAKERS

The judges who participate in a standard-setting session do not have the final decision regarding what standard should be set. They serve in an advisory capacity to the board that has the official responsibility and authority to set the standard. As Popham (1987) pointed out, "Those who set standards for high-stakes tests are a decisively endangered species" (p. 77). However, as he goes on to say, the dangers are not extinction, but, rather, political and legal. To minimize the political and legal dangers, policymakers may need guidance on how to proceed in setting the standards.

Someone (typically an executive director) needs to prepare the board members for their important decision. That preparation may vary considerably from board to board depending on the sophistication of the members. However, all boards should understand that all standards are fundamentally judgmental and that no *technical* mechanism can be used to prevent judgments *or* to prove that any set of judgments are the "correct" ones. The board must understand the importance of the standard, the implications of the decisions regarding the standard, the nature and quality of the examination, the recommendations of the judges, the judges' qualifications and training, and, if available, the impact of the standard based on field test results, or *preferably* on live test results.

After an orientation with the board members regarding their responsibilities, the recommendations of the judges should be thoroughly communicated to them. This should include the mean, median, range, and standard deviation of the judges' standards, as well as information on the intrajudge consistency (if available). The board members should specifically consider the relative costs of false positives and false negatives. Finally, after setting the official standard, the board needs to consider how the standards should be made public.

SUMMARY

Concerns Related to Discussions About Standard-Setting Methodologies

This paper began with a discussion of some general concerns related to the discussions on standard-setting methodologies. These included the following notions:

1. We should separate policy/political views from technical considerations.
2. Context matters in choosing methods.
3. Item type may impact standard-setting methodologies.
4. We should not overgeneralize about "best methods" across contexts.

5. We should not ignore previous research and draw conclusions on a single study.

6. We should separate views about the quality of the item pool from evaluation of the standard-setting methodology.

7. Adequacy of item exemplars in communicating standards is separable from the "correctness" of the standards.

Pre-1992 Literature Review

Although the literature on standard setting is inconclusive on many points, there seems to be agreement that standard setting is a judgmental process and that there is no "right answer" regarding what standard should be set.

One common classification of approaches to standard setting is to divide them into two broad categories: state and continuum (Berk, 1986). State models were not considered in this paper because they make the unrealistic assumption that the traits in question are truly dichotomous in nature.

Another level of classification differentiates between procedures that set standards and those that adjust standards. Those that adjust standards assume some standard has already been identified and were not reviewed in this paper. The classification used here was examinee-centered models, test-centered models, and compromise models.

The major approaches in the test-centered category are the Angoff (1971) Ebel (1979), and Nedelsky (1954) models (or variations thereof). A large number of studies have compared these models. In considering these studies, several points need to be kept in mind. First, different models should not be expected to lead to the same standard. Second, because there is no known true standard, the standards derived from different methods cannot be compared with respect to degrees of correctness. Third, the studies are all based within a particular context, and there is no particular reason to believe that we can generalize from one context to another. Fourth, there are many variations of the three methods, and a research study using one variation may not generalize to another variation. Nevertheless, some general (and sometimes tentative) conclusions can be drawn from the literature.

Typically the Nedelsky method yields the lowest standards, and the Ebel method yields the highest standards, with the Angoff method in the middle. The Angoff procedure tends to produce the more consistent (reliable) standards. Intrajudge consistency has generally been found to be higher for Angoff's method. Further, there are not as many logical/psychological arguments against the Angoff method.

The Angoff method is the most popular of the three methods, and my recommendation is that it should be considered the preferred test-centered model. Logic suggests, however, that judgments made using this approach need to be tempered by reality, and there is some evidence showing that the Angoff model works best when there is feedback to the judges regarding what empirical data indicate regarding the difficulty of the items and the distribution of the examination scores. Thus, it is important to consider empirical data.

Jaeger (1978) first suggested combining a test-centered model with an iterative procedure that included some empirical information. Currently, a very common iteration procedure is for judges to rate each item via the Angoff approach as an initial step. Then in subsequent iterations they are given information about how the other judges responded, about the difficulty of the items, and about the impact (passing rate) of various proposed standards.

Another set of approaches using empirical data is referred to as compromise models. The three best known compromise models are the Hofstee (1983), Beuk (1984), and De Gruijter (1985) models. The Hofstee and Beuk models are the most popular. Both require the judges to set a passing score directly and to incorporate those judgments into a model that includes knowledge of actual examinee performance. Both models are relatively easy to use, and some studies suggest they result in similar standards. The choice between them is generally a matter of individual taste, although the Beuk method makes a questionable assumption regarding the meaning of the judges' standard deviations and the Hofstee model may require an extrapolation of a line in order to set a standard.

Recent Approaches to Standard Setting

Because most of the pre-1992 research was done on assessments where the item scoring was dichotomous and because much current assessment in education uses performance (constructed-response) assessment with polychotomous scoring, much current research has focused on setting standards for such assessments. The National Academy, ACT/NAGB, NBPTS, various states, and some noneducational agencies have all conducted research on new or modified methodologies. Many methods look promising. None have a sufficient research base to be considered as established as the judgmental-empirical methods commonly used for tests composed of dichotomous items.

Summary of the Literature Review

A multitude of standard-setting methodologies exist. Historically, preference has been shown for judgmental-empirical approaches. The modified (iterative with data presented) Angoff method has been the most popular. There has been some recent criticism of this model, and some question its appropriateness for polychotomously scored items (tasks). Recent research has concentrated on methods for polychotomous items. This paper has proposed 19 areas of agreement and 5 areas of disagreement among the "experts" on standard-setting methodologies. Slightly abbreviated versions of these areas are stated below.

Agreements

1. Standards are judgmental; there is no right answer as to where a standard should be set.
2. Context matters.
3. Test-centered methods, if used, should be supplemented by information on item statistics and impact data.
4. Impact data should be based on real test data, not on field test data.
5. The Angoff is the preferred test-centered method for dichotomous items.

6. Test-centered methods require judges to see the items on the assessment to set standards.

7. An item-by-item approach forces item scrutiny.

8. Examinee-centered methods require judges to determine each test taker's knowledge and skills independent of the test score.

9. This independent judgment is either hard or impossible to do well.

10. The contrasting-groups method is the most popular of the examinee-centered methods.

11. Judges must have some understanding of what borderline means.

12. This understanding should not be changed after the standards are set.

13. Judges must understand the methodology but do not have to understand its math.

14. Setting standards for polychotomous items is not as well researched as it is for dichotomous items.

15. Judges should see a sample of examinee papers prior to setting standards on performance assessments.

16. Standards should not be based on *only* a consideration of scoring rubrics.

17. Subjective scoring of performance is corruptible, and scoring standards may change after judges know what the passing standard is.

18. Conjunctive models prohibit totaling task scores.

19. Judges cannot readily understand the effects of conjunctive scoring on passing rates.

Disagreements

There is strong disagreement among scholars about the truthfulness of the following propositions. My opinion is given in parentheses after the proposition.

1. Empirical findings regarding what examinees can do, independent of the assessment, is evidence regarding the correctness of judges' beliefs about what examinees should be able to do. (This depends on whether "correctness" refers to the correct reporting of the belief or to a correct belief.)

2. Variance across judges' ratings is a relevant criteria in judging the adequacy of the method. (Such variance is relevant to determining what sample size is wanted but not to method adequacy.)

BEST COPY AVAILABLE

3. Item-by-item approaches are fatally flawed. (Twenty years of research has shown this view to be wrong.)

4. It is reasonable to have nonexperts make judgments about standards. (Most experts would probably disagree with this statement, unless they are wearing a policymaker's hat.)

5. If we cannot have perfect standards, let us not have them. (I disagree. We still must often make categorical decisions. These are made on either implicit or explicit standards. Explicit standards are better.)

Other Considerations

Setting a standard involves many considerations in addition to the particular standard-setting method that is used. If judges are to be used, it must be determined how to identify the appropriate potential set of judges, how to choose from that population, and how many judges need to be chosen. Judges should be expert in content knowledge. Judges should be demographically representative. There should be a sufficient number of judges to provide a reasonably small standard error of the mean (some literature suggests from 13 to 21 judges).

Judges must be thoroughly trained, and their judgmental process must be monitored with feedback given to them regarding intrajudge consistency and interjudge consensus.

There are several ways to aggregate the individual judges' standards (e.g., mean, median, or trimmed mean). It is suggested that the individual judge's standards and the various aggregations should be made available to the board.

Preparing Policymakers

Someone needs to prepare the Board to make the final decision regarding the standard. The Board must understand the importance of the standards, the implications of the decisions regarding the standard, the nature and quality of the examination, the recommendations of the judges, the judges' qualifications and training, and, if available, the impact of the standard. Board members should probably be told specifically to consider the relative costs of false positives and false negatives. Finally, after setting the official standard, the Board needs to consider how this standard should be made public.

References

American College Testing. (1993). *Setting achievement levels on the 1992 National Assessment of Educational Progress in mathematics, reading, and writing: A technical report on reliability and validity*. Iowa City: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Arrasmith, D. G., & Hambleton, R. K. (1988). *Steps for Setting Standards with the Angoff Method.* Washington, DC: EDRS.

Behuniak, P., Jr., Archambault, F. X., & Gable, R. K. (1982). Angoff and Nedelsky standard setting procedures; Implications for the validity of proficiency test score interpretation. *Educational and Psychological Measurement, 42,* 247-255.

Berk, R. A. (1984). *Screening and diagnosis of children with learning disabilities*. Springfield, IL: Charles C. Thomas.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56,* 137-172.

Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement, 21,* 147-152.

Bowers, J. J., & Shindoll, R. R. (1989). *A comparison of the Angoff, Beuk, and Hofstee methods for setting a passing score.* Iowa City, IA: American College Testing.

Brennan, R. D., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement, 4,* 219-240.

Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement, 27,* 145-163.

Cizek, G. J. (1993). *Reactions to National Academy of Education report, setting performance standards for student achievement*. Washington, DC: National Assessment Governing Board.

Cizek, G. J. (1991). *An investigation into one alternative to the group process procedure for setting performance standards on a medical specialty examination*. Upublished dissertatioin, Michigan State University.

Colton, D. A., & Hecht, J. T. (1981, April). *A preliminary report on a study of three techniques for setting minimum passing scores.* Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles.

Cope, R. T. (1987, April). *A generalizability study of the Angoff method applied to setting cutoff scores of professional certification tests.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. *Journal of Educational Measurement, 21,* 113-129.

De Gruijter, D. N. M. (1985). Compromise models for establishing examination standards. *Journal of Educational Measurement, 22,* 263-269.

Ebel, R. L. (1979). *Essentials of educational measurement* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor and Department of Justice. (1978, August 25). Uniform guidelines on employee selection procedures. Fed. Reg., 43(166), 38290-38315.

Friedman, C. B., & Ho, K. T. (1990, April). *Interjudge consensus and intrajudge consistency: Is it possible to have both in standard setting?* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston.

Garrido, M., & Payne, D. A. (1987, April). *An experimental study of the effect of judges' knowledge of item data on two forms of the Angoff standard setting method.* Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Gross, L. J. (1985). Setting cutoff scores on credentialing examinations: A refinement in the Nedelsky procedure. *Evaluation and the Health Professions, 8,* 469-493.

Harker, J. K., & Cope, R. T. (1988, April). *The effect of several variables on judgmentally-obtained cut scores.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Hofstee, W. (1973). *Een alternatief voor normhandhoving bij toetsen. Netherlands Tijdschrift von de Psychologie, 28,* 215-227.

Hofstee, W. (1983). The case for compromise in educational selection and grading. In S.B. Anderson and J.S. Helmick (Eds.), *On educational testing* (pp. 109-127). San Francisco: Jossey-Bass.

Jaeger, R. M. (1978). *A proposal for setting a standard on the North Carolina high school competency test.* Paper presented at the spring meeting of the North Carolina Association for Research in Education, Chapel Hill.

Jaeger, R. M. (1988). Use and effect of caution indices in detecting aberrant patterns of standard-setting judgments. *Applied Measurement in Education, 1,* 17-31.

Jaeger, R. M. (1989). Certification of student competence. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). New York: American Council on Education/Macmillan.

Jaeger, R. M. (1990a). Establishing standards for teacher certification tests. *Educational Measurement: Issues and Practice, 9,* 15-20.

Jaeger, R. M. (1990b). Setting standards on teacher certification tests. In J. Millman and L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation* (pp. 295-321). Newbury Park, CA: Sage.

Jaeger, R. M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice, 10*(2), 3-6, 10, 14.

Jaeger, R. M. (1994, April). *Setting performance standards through two-stage judgmental policy capturing*. Paper presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, New Orleans.

Jones, J. P. (1987). *The effects of the job status of judges and the presence of item statistics on the passing scores set by three pooled-judgment methods.* Doctoral dissertation, Columbia University. (University Microfilms No. ADG87-24043, 9000)

Julian, E. R. (1993, April). *Standard setting on performance examinations*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Kahl, S. R., Crockett, T. J., & DePascale, C. A. (1994, June). *Using actual student work to determine cut scores for proficiency levels: New methods for new tests*. Paper presented at the National Conference on Large-Scale Assessment sponsored by the Council of Chief State School Officers, Albuquerque, NM.

Kane, M. (1993). *Comments on the NAE evaluation of the NAGB achievement Levels*. Washington, DC: National Assessment Governing Board.

Kane, M. (1994). *Criterion bias in examinee-centered standard setting*. Washington, DC: National Assessment Governing Board. [Xerox].

Klein, L. W. (1984, April). *Practical considerations in the design of standard setting studies in health occupations*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Klein, S. P. (1991, April). *Setting pass/fail standards on constructed response licensing tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Korte, R. C. (1987). *A comparison of four methods for setting C min. and C max. in Hofstee's compromise standards setting model*. Doctoral dissertation, Michigan State University. (University Microfilms No. ADG88-01830. 9000)

Linn, R. L. (1984, April). *Standards for validity in licensure testing.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Linn, R. L., Koretz, D. M., Baker, E. L., & Burstein, L. (1991). *The validity and credibility of the achievement levels for the 1990 National Assessment of Educational Progress in mathematics* (CSE Rep. 330). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Luecht, R. M. (1993, April). *Using IRT to improve the standard setting process for dichotomous and polytomous items.* Paper presented at the annual meeting of the National Council of Measurement in Education, Atlanta.

McLaughlin, D. H. (1993, May). *Rated achievement levels of completed NAEP mathematics booklets.* Paper prepared as part of the National Academy of Education's evaluation of the 1992 trial state assessment.

McLaughlin, D. H., Dubois, P., Eaton, M., Erlich, D., Stancavage, F. B., O'Donnell, C., Yu, J., DeStefano, L., Pearson, P.D., Bottomley, D., Bullock, C. A., Henson, M., & Rucinski, C. (1993, (DRAFT) June). *Teachers' and researchers' ratings of student performance and NAEP mathematics and reading achievement levels.* Paper prepared as part of the National Academy of Education's evaluation of the 1992 trial state assessment.

Mehrens, W. A. (1990). Combining evaluation data from multiple sources. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 322-334). Newbury Park, CA: Sage.

Melican, G. J., Mills, C. N., & Plake, B. S. (1989). Accuracy of item performance predictions based on the Nedelsky standard setting method. *Educational and Psychological Measurement, 49,* 467-478.

Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. *Review of Educational Research, 46,* 133-158.

Meskauskas, J. A. (1983, April). *Standard-setting: State of the art, future prospects.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Meskauskas, J. A. (1986). Setting standards for credentialing examinations. An update. *Evaluation and the Health Professions, 9,* 187-203.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Mills, C. N., & Melican, G. J. (1987). *A preliminary investigation of three compromise methods for establishing cut-off scores.* Princeton, NJ: Educational Testing Service.

Mills, C. N., & Melican, G. J. (1988). Estimating and adjusting cutoff scores: Features of selected methods. *Applied Measurement in Education, 1,* 261-275.

Mills, C. N., Melican, G. J., & Ahluwalia, N. T. (1991). Defining minimal competence. *Educational Measurement: Issues and Practice, 10*(2), 7-10.

Mullins, M., & Green, D. R. (1994). In search of truth and the perfect standard-setting method: Is the Angoff procedure the best available for credentialing? *Clear Exam Review, 5*(1), 21-24.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement,* 14, 3-19.

Norcini, J. J., Shea, J. A., & Kanya, D. T. (1988). The effect of various factors on standard setting. *Journal of Educational Measurement, 25,* 57-65.

Norcini, J. J., Stillman, P. L., Sutnick, A. I., Friedman, M., Heley, H., Regan, M. B., Williams, R. (1993). Scoring and standard setting with standardized patients. *Evaluation and the Health Professions, 16,* 322-332.

Plake, B. S. (1994, April). *An integration and reprise: What we think we have learned*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Poggio, J. P. (1994). *Practical considerations when setting test standards: A look at the process used in Kansas*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Poggio. J. P., & Glasnapp, D. R. (1994, April). *A method for setting multi-level performance standards on objective or constructed response tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Popham, W. J. (1987). Preparing policy makers for standard setting on high-stakes tests. *Educational Evaluation and Policy Analysis, 9,* 77-82.

Putnam, S. E., Pence, P., & Jaeger, R. M. (1994, April). *A multi-stage dominant profile method for setting standards on complex performance assessments*. Presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, New Orleans.

Pyburn, K. M., Jr. (1994, April). *Legal challenges to licensing examinations*. Paper presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, New Orleans.

Rebell, M. A. (1990). Legal issues concerning teacher evaluation. In J. Millman and L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation* (pp. 337-355). Newbury Park, CA: Sage.

Reckase, M. D. (1994, April). *Standard setting on performance assessments: A comparison between the paper selection method and the contrasting groups method*. Paper presented at the National Conference on Large-Scale Assessment sponsored by the Council of Chief State School Officers, Albuquerque, NM.

Reid, J. B. (1991). Training judges to generate standard-setting data. *Educational Measurement: Issues and Practice, 10*(2), 11-14.

Reilly, R. R., Zink, D. L., & Israelski, W. E. (1984). Comparison of direct and indirect methods for setting minimum passing scores. *Applied Psychological Measurement, 8,* 421-429.

Richardson v. Lamar County Board of Education et al., No. 87-T-568-N (1989); U.S. C. App. 11th Cir. Nos. 90-7002, 90-7336, July 17, 1991.

Rock, D. A., Davis, E. L., & Werts, C. (1980, June). *An empirical comparison of judgmental approaches to standard-setting procedures* (ETS Res. Rep.). Princeton, NJ: Educational Testing Service.

Rothman, A., Poldre, P., Cohen, R., & Ross, J. (n.d.). *Standard setting in a multiple station test of clinical skills* [Manuscript].

Shepard, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 169-198). Baltimore: John Hopkins University Press.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education, 19* (pp. 405-450). Washington, DC: American Educational Research Association.

Shepard, L. A., Glaser, R., Linn, R. L., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement*. A report of the National Academy of Education Panel on the Evaluation of the NAEP trial state assessment: An evaluation of the 1992 achievement levels. Stanford, CA: Stanford University, National Academy of Education.

Skakun, E. N., & Kling, S. (1980). Comparability of methods for setting standards. *Journal of Educational Measurement, 17,* 229-235.

Smith, I. L., & Greenberg, S. (1990, April). *Strategies for making cut score determinations on a performance-based, observational test: Design versus implementation.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Smith, R. L., & Smith, J. K. (1988). Differential use of item information by judges using the Angoff and Nedelsky procedures. *Journal of Educational Measurement, 25,* 259-274.

Stufflebeam, D. L., Jaeger, R. M., & Scriven, M. (1991). *Summative evaluation of the National Assessment Governing Board's inaugural 1990-91 effort to set achievement levels on the National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board, August 1991.

Tyler v. Vickery, 517 F.2d 1089 (5th Cir. 1975), cert. denied, 426 U.S. 940 (1976).

van der Linden, W. J. (1982). A latent trait method for determining intrajudge inconsistencies in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement, 19,* 295-308.

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow and D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 75-107). Hillsdale, NJ: Erlbaum.

Zieky, M. J. (1989). Methods of setting standards of performance on criterion referenced tests. *Studies in Education Evaluation, 15,* 335-338.

Concerns Related to Discussions About Standard-Setting Methodologies

Summary of Break-out Session[1]

The session began with a discussion of *the best method for setting cut scores on tests that use both dichotomous (multiple choice) and polytomous items (responses which can be scored as ordered categories)*. Participants first shared their experiences with one or the other method. One participant cited the cases of Philadelphia and New Jersey with respect to polytomous items. He reported that in the 1992 National Assessment of Educational Progress (NAEP), Philadelphia used *the paper-selection method,* a random selection of papers that exemplified each of the scores. This led to the identification of cut scores for polytomous items that were considerably higher than those set for the dichotomous items. In New Jersey, *a polytomous approach* (i.e., identifying and scoring papers above and below the cut score) was used. This resulted in cut scores that were considerably lower than that for the dichotomous method. This participant hypothesized that the difference may have been in the way the algorithms were assembled based upon the contrasting-groups method. He added that the two approaches might have been measuring different constructs. It was suggested that it is necessary to investigate further, first, if these two methods really work, and, second, whether these two methods can be combined.

Another participant shared his use of *a modified Angoff method* and *a traditional contrasting-groups approach.* He indicated that the modified Angoff method produced higher standards, as opposed to the contrasting-groups approach, in which most were judged as "not proficient." It was suggested that the differences existed because (a) in the Angoff method, more intensive training was provided to judges, and (b) the contrasting-groups approach used the actual scores of students.

It was noted that different approaches result in different outcomes not because the methods are faulty, but rather because each method scores in terms of different standards. Another participant noted that in the contrasting-groups approach, there is no single cut score but a distribution of scores; he suggested that if a point of reference is set high enough, then the outcome will come out closer to the Angoff method. He explained that the outcomes from both methods result from *different weighting of two different classification areas*. Thus the classification error between the groups must be considered. He urged no further comparison.

Mehrens cited *another approach* as used in the state bar examination of *California* where judges looked at previously scored papers, sorted them into two groups, and then took the differences in the scores given by the two groups (the previous and the current judges). The result became the cut score for the polytomous items. Then, the percent failure for the polytomous items was used to set the cut score for the dichotomous portion of the test; the scores for the polytomous items were considered more reliable.

Mehrens also noted the dilemma of measuring a construct, which the Item Response Theory (IRT) model suggests is unidimensional, but for which one has the information function of the polytomous items and the dichotomous items. The dichotomy here was to ignore that information in setting the standards while remaining aware of its importance in setting standards.

---

[1] This is a summary of the break-out discussions for William Mehrens' presentation. The session was facilitated by Luz Bay (American College Testing) and recorded by Sal Corallo (National Center for Education Statistics).

One participant suggested that the accuracy of the judges' probability estimates was just as important as the possible error resulting from the different methodologies. He questioned whether *judges can give accurate probability estimates as required in the Angoff method*. Mehrens suggested that since the correct cut score is not known, it is difficult to address the accuracy of the judges. However, he indicated that we can recognize cut scores that are on the extremes.

One participant argued on behalf of the *Nedelsky approach*. He proposed that this think-aloud protocol provides some agreement between the elimination of the distractors in multiple-choice items on the part of the students and the elimination of the distractors on the part of the judges. He indicated that he is trying this approach with a small sample. The group posed questions related to the representativeness of the students, the information that judges were told about the students, and how the data were summarized. One recommendation was that the judges' opinions at least should be validated by some outside criteria.

After some additional discussion, the group proposed that no one method was better than another, but that there may be occasions when one method is the best for the context. One participant suggested that if there were uneasiness with the results of one method, it might be better to replicate the process using the same approach to obtain consistency rather than using a different method.

The discussion then focused on the use of external standards. There were two main positions presented. In the first, it was argued that the nature of standards requires item-judgment study. In the second, it was suggested that NAEP levels are at least partly predictive and, as such, should be open to empirical validation. This latter position was contested; it was suggested that the NAEP standards were never meant to be predictive or prescriptive. Mehrens indicated that it was a mistake to use the standards in a predictive fashion, because, as such, they beg for predictive validity evidence. He suggested that the uses of the NAEP standards must be stated clearly or proof of predictive validity must be provided.

At this point in the discussion, there was agreement in the group that there should be external validation of standards, but there was no agreement on the type of external evidence that should be gathered other than that it should be based on the purpose of the assessment.

The group discussed briefly the process of standard setting. One participant questioned if judges need to view examples of actual performance before setting standards. It was suggested that judges should be given impact data. The discussion then focused on the public's understanding of standard setting and of the NAEP reports generated thus far. It was suggested that *a norm-based approach* to standard setting may better inform the public than a criterion-based approach. Mehrens referenced the case of North Carolina, which uses what he called a *norm-referenced approach with contrasting groups* based on teachers' judgments.

The group then discussed the composition of the panel for standard setting and the panelists themselves. Some argued that members of the public should be part of the panels so that they can contribute the perspective of the consumer. There was a suggestion that the characteristics of the panelists ought to be different for the different levels of the examination. This suggestion did not find much support.

Participants inquired about the importance of the judges' knowledge of the students' opportunity to learn. It was explained that it would be of little importance if the purpose of the assessment was to raise the level of learning. Additional questions were then asked about the purpose and consequences of the test. It was noted that the contexts differ across the grades and so the purposes differ; cut scores have different purposes at different grades. It was suggested that the pure norm-referenced approach can be used for standard setting, as well as IRT models. One participant indicated that in this context, it is necessary to determine what students below the cut scores need to have in terms of remediation.

The group discussed the representativeness of the panel and the elimination of judges. In the discussion participants explored the training procedures for judges and methods for dealing with judges whose rankings are outliers (i.e., far above or below the average rankings determined by the panel). It was noted that it is possible to get consensus across different judges (as contrasted to agreement); however, what is more important is to ensure that judges are aware of what they are required to do.

The group ended its discussion with an examination of how a borderline student is defined and the reasons for determining a borderline category. Participants indicated that it is difficult to define the borderline student prior to tests. They determined that there is a need for further information about reasons for determining a borderline category.

Standard Setting From the Perspective of Generalizability Theory[1]

Robert L. Brennan

Professor of Educational Measurement, Director of Iowa Testing Programs

ABSTRACT

In education currently, the term "standards" has numerous meanings. In this paper, the word standards *is used to refer to* standards *for standard-setting procedures. Otherwise, "standards" refers either to standard-setting processes or to outcomes of such processes. A substantial part of the audience for this paper includes persons with considerable experience in setting or evaluating National Assessment of Educational Progress (NAEP) achievement levels, which are certainly standards. While it is clearly intended that this paper address issues of generalizability in establishing achievement levels, the focus on standards considered here is broader than just achievement levels.*

This paper is organized into four principal parts. The first part provides a conceptual framework for considering standard setting, broadly conceived, in the context of generalizability theory. The second part considers the role of various facets (including judges, items, occasions, and methods) in characterizing the dependability of the outcomes of a standard-setting process. The third part treats some collateral issues, including different perspectives on dimensionality and the role of score scales in standard setting. Finally, the fourth part suggests some particular standards *for standard setting. This paper deals largely with reliability-related issues, but it also considers aspects of validity.*

In education currently, the term "standards" has numerous meanings. Politicians, the press, and the public often view standards as goals that some agency or body declares desirable. Sometimes these goals are essentially content standards; at other times they are performance standards. Usually, such goals are stated rather generally. To operationalize them, a standard-setting study is often undertaken that uses a particular method to arrive at outcomes, or standards, that effectively serve as advice to policy/decision makers responsible for establishing standards. Ultimately, this leads to the question, "What *standards* should apply to the process of standard setting to ensure that the outcomes (or standards) are defensible?"

The multiplicity of connotations for the term "standards" is at best confusing. Further, any of these connotations can be, and often are, loaded with excess meaning. After all, many people are psychologically uncomfortable arguing against standards. Yet standards are not "truth"; they are not "right"; and the existence of different standards is not necessarily a contradiction. Although standards lack "truth value," they can have "worth" for particular purposes in particular contexts. In the abstract, however, "standards" do not have worth.

When standards get encapsulated in labels, the possibility of misinterpretation becomes a serious threat. For example, characterizing as "advanced" examinees who score in a particular range on a test does not necessarily mean that they all will be successful in a subsequent activity. What "advanced" means can only be judged by examining the agreed-upon definition of "advanced" as well as any collateral information.

---

[1] The author gratefully acknowledges many helpful discussions with Michael T. Kane that influenced several parts of this paper.

The standard-setting issues considered in this paper are treated largely through generalizability theory. However, it is not the purpose of this paper to provide an integrated treatment of the theory. A full explication of the concepts and the methods of generalizability theory is provided by Cronbach, Gleser, Nanda, and Rajaratnam (1972). Briefer treatments are provided by Brennan (1992a), Feldt and Brennan (1989), and Shavelson and Webb (1991), and Brennan (1992b) provides an introduction.

A CONCEPTUAL FRAMEWORK

This section provides a conceptual framework that forms a basis for considering the generalizability of results for a standard-setting process. This framework has three components: (a) the outcomes of the process; (b) the standard-setting methods that might be used; and (c) the notion of "replication" and how it relates to random errors of measurement.

Outcomes

The standard-setting processes considered here have, in theory, three outcomes: (a) performance standards or definitions; (b) cut scores; and (c) exemplars (e.g., items or booklets). The distinction between performance standards and cut scores has been treated in considerable detail by Kane (1994). (Also, the "conceptual" standards of Livingston, 1994, are closely related to the performance standards of Kane, 1994.) In subsequent discussions, performance standards will usually be referred to as definitions. This is done in part to minimize confusion with respect to multiple connotations of the word "standards." In addition, the term "definitions" is usually used in conjunction with the achievement-level-setting processes of the National Assessment of Educational Progress (NAEP). In all cases, however, the word "definitions" as used here is closely associated with Kane's notion of performance standards.

In practice, definitions, cut scores, and exemplars may not all be present, or there may be other relevant, hidden elements. For example, definitions are sometimes specified a priori. Also, a paucity of available items may seriously restrict the pool of items from which raters can choose exemplars--a restriction that casts doubt on how "exemplary" such items truly are. In addition, the cut scores resulting from raters' judgments are usually raw scores for a set of items or test booklet, as opposed to scale scores that are the ultimate basis for standards. Finally, it is rare for judges to have the ultimate authority to establish standards, and their recommendations may not be accepted by those who do have authority. However, this paper assumes that definitions, cut scores, and exemplar items are all potential outcomes of the standard-setting process, although this paper emphasizes cut scores.

Methods

I have made no attempt to catalog the many standard-setting methods currently available (see Jaeger, 1989) or to characterize them in a tight conceptual framework (see Kane, 1994). Rather, I have differentiated methods in terms of whether they are (a) item/test centered; (b) examinee centered; or (c) scale centered. In this scheme, the Angoff (1971, modified or not), Nedelsky (1954), and various whole-booklet procedures are item/test-centered methods. The borderline and contrasting-groups procedures are examinee-centered methods. And the anchoring procedure that has been used in NAEP is a scale-centered method. I recognize that some might quarrel with characterizing anchoring as a standard-setting process, and I do not give great attention to anchoring in this paper. However, it seems sensible to include anchoring in general discussions.

For present purposes, a full understanding of these methods is unnecessary, but an appreciation of some differences is useful. Although it may be an oversimplification, both the item/test and examinee methods start with definitions and then arrive at cut scores. In contrast, the scale-centered anchoring procedure starts with a scale score and then produces a definition. It is virtually certain, therefore, that replication of an anchoring procedure will lead to the same scale score but different definitions, whereas replication of any of the other procedures will produce different scale scores and different definitions (unless definitions were specified a priori).

The item/test- and examinee-centered methods differ in the focal point of raters' efforts. For item/test-centered methods, the focal point is the performance of hypothetical examinees on actual items that represent a prespecified content framework. For examinee-centered methods, the focal point is the estimated level of achievement of actual examinees known to the judges, with the congruence between examinee achievement levels and definitions being an additional matter of judgment. In short, item/test-centered methods are grounded in the reality of actual items representing the framework, whereas examinee-centered methods are grounded in the reality of actual examinees known to the judges. This characterization of methods is admittedly an oversimplification, but it does help to clarify what is fixed and what is variable from the perspective of the judges.

The intent is that most of the basic principles discussed in this paper should apply to all these types of methods. Of course, the differences among the types of methods render some discussions less applicable to certain methods than to others. This dissonance, when it occurs, is much more tolerable than tailoring discussion to one specific method, when one of the goals is to suggest technical *standards* for standard setting. For expository purposes, however, it is often convenient to couch discussions in terms of item/test-centered methods and let readers, at their discretion, translate the discussion to other types of methods when it is sensible to do so. Occasionally, footnotes are used to aid readers in such translations.

Replications

To document the trustworthiness of the results of a standard-setting study, it is necessary to answer the question, "How variable would results be if the process were replicated?" Figure 1 illustrates how the results of multiple replications might be documented, along with the "average" results. Obviously, Figure 1 represents a hypothetical situation because it is rarely the case that practical constraints permit obtaining replications. Also, it is by no means unambiguously clear how an investigator would reconcile differences (over replications) among definitions to obtain "average" definitions for operational use. A comparable problem exists for exemplars.

Nonetheless, the question of comparability across replications is central to any meaningful understanding of the trustworthiness of the results of a standard-setting study. Because it is usually impossible to perform replications in the strict sense of the term, relatively strong statistical assumptions (or perhaps a minireplication) are typically required. In any case, even if replications cannot be obtained, a central issue that must be addressed is, "What would constitute a replication?" Or, in the terminology of generalizability theory, "What are the facets in the universe?" Generalizability theory

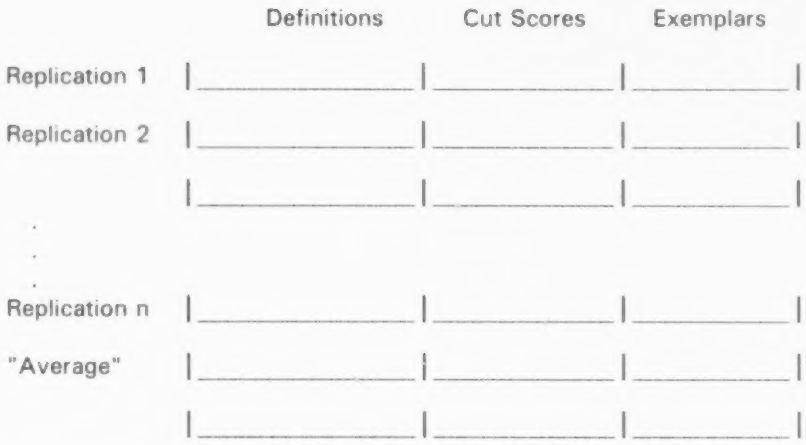

Definitions
Cut Scores
Exemplars
Replication 1
Replication 2
Replication n
"Average"

Figure 1. Standard-setting replications.

per se does not provide an answer to this question. It is the investigator's or policymaker's responsibility to specify the facets over which generalization is to be considered.

Suppose an investigator, Smith, decided that the facets will be raters, items, and occasions. For a universe characterized by these three facets, each replication would involve a different set of raters, a different set of items, and a different occasion. A different investigator, Jones, might decide to use the same items in every replication. If so, Jones' universe would be narrower than Smith's, and it is almost certain that a standard error of measurement based on Jones' narrower universe would be smaller than for Smith's universe. This does not mean that Jones' universe is better than Smith's; it merely means that the two investigators have a different conceptualization of the universe.

This thought experiment also illustrates that, strictly speaking, there is no such thing as *the* standard error of measurement. There are numerous possible standard errors of measurement corresponding to different universes and designs. Indeed, standard errors of measurement often can be made arbitrarily large or small by broadening or narrowing the universe. Therefore, statements about standard errors of measurement should not be judged in the abstract but should be interpreted relative to a clear specification of the universe. For a standard-setting study based on item/test-centered or examinee-centered methods, standard errors of measurement for the cut scores are statistics of primary interest because they provide a direct answer to the question, "How variable would cut scores be if the process were replicated?" At present, there is no statistic in general use that characterizes the variability in definitions and/or exemplars that might occur over replications. However, if the standard-setting study involves developing definitions and choosing exemplars, then variability among definitions and/or exemplars should be described somehow as part of a full characterization of the generalizability of the process.

FACETS AND GENERALIZABILITY

As has been noted, it is the investigator's or policymaker's responsibility to specify the facets in the universe. It seems sensible that the universe should take into account variability attributable to at least judges, items, and occasions.[2,3] (Some would argue that method also should be a facet in the universe.) All of these facets are considered in this section, although they are not always treated in a strictly independent manner. A hypothetical design illustrates how variability attributable to two of these facets might contribute to measurement error.

A Hypothetical *D* Study Design

In generalizability theory, the data collection design that is actually used to make decisions is called a *D* study. Here, for expository purposes, a particular design is introduced that involves potential sources of variability attributable primarily to the judge and item facets. Doing so facilitates subsequent discussion of issues associated with facets, provides an illustration of a particular standard error of measurement, and leads to some reliability/validity distinctions. Although the design to be discussed

---

[2] Variability over items is clearly an issue for item-centered methods, but it is also an issue for examinee-centered and scale-centered methods. For example, results for the contrasting-groups method may vary depending on which items (item sets or test booklets) are taken by examinees who are assigned to the groups. Also, the anchoring method usually gives quite different definitions of scale score points for different forms of a test.

[3] For an examinee-centered method, it would seem sensible that examinees be a facet, too.

is useful for illustrative purposes, it is not prescriptive because it does not reflect all the complexities of an actual standard-setting study.

Suppose the universe of potential judges (j) is subdivided into $n_s = 3$ strata (s), and the universe of potential tasks or items (i) is subdivided into $n_t = 2$ types (t) of items. Suppose further that any replication consists of sampling n_j judges from each of the strata for judges and n_i items from each of the item types. (An equal number of judges for all strata and an equal number of items of both types are unlikely, of course; but this and other simplifications are necessary here to keep complexity within reasonable bounds.) Finally, assume that in principle, any judge could evaluate any item. In this case, judges within strata and items within type are random effects; strata and item types are fixed effects.

For this design, the standard error *(SE)* of measurement for mean scores over all $3n_j$ judges and $2n_i$ items is

$$SE = \left[\frac{\sigma^2(j:s)}{3n_j} + \frac{\sigma^2(i:t)}{2n_i} + \frac{\sigma^2(ji:st)}{6n_j n_i} + \frac{\sigma^2(si:t)}{6n_i} + \frac{\sigma^2(tj:s)}{6n_j} \right]^{1/2}, \qquad (1)$$

where the variance components (σ^2) are:

$\sigma^2(j:s)$ = the average (over strata) of the variances among judges in a stratum;

$\sigma^2(i:t)$ = the average (over item types) of the variances of the items of a particular type;

$\sigma^2(ji:st)$ = the average of the interactions of judges and items within each stratum-type combination;

$\sigma^2(si:t)$ = the average, over item types, of the interactions of strata and items; and

$\sigma^2(tj:s)$ = the average, over strata, of the interactions of item types and judges.

For those versed in generalizability theory, *SE* in Equation 1 is an absolute (Δ-type) standard error, as opposed to a relative (δ-type) standard error. When performance relative to a fixed standard is the primary concern, as it is in standard-setting contexts, the error of interest is absolute error.

Equation 1 is somewhat more general than its form might suggest. In particular, *i* need not stand for a single task or item. Rather, *i* could represent a set of items or, perhaps, a test booklet. If so, then the variance components involving *i* would be variances associated with booklets. For example, $\sigma^2(i:t)$ would be the variance component for booklets of a particular type, and n_i would be the number of booklets.[4]

There are at least four important points to make about *SE* in Equation 1. First, *SE* is for *mean* scores over *both* items and judges; it is not a standard error of measurement for individual judges. Second,

---

[4] Under some circumstances, equating could have the effect of setting $\sigma_2(i:t)$ to zero in Equation 1. A similar result might occur for one or more of the variance components involving interactions of *i* with other facets. Before assuming that such variance components are zero, however, an investigator should be certain that the equating process would indeed cause these variance components to be zero. For this to occur, the equating transformation applied to examinee responses would have to be applicable also to judges' ratings. Rarely would this seem to be a sensible assumption.

SE clearly becomes smaller as n_j and/or n_i increase. In this sense, having more judges rate more items results in improvement in measurement precision. Third, the error taken into account by *SE* is random error, and, in that sense, *SE* is an indicator of unreliability. Fourth, variances attributable to overall mean differences in types of judges [$\sigma^2(s)$], types of items [$\sigma^2(t)$], and their interaction [$\sigma^2(st)$] do not contribute to random error of measurement for mean scores. This implies, for example, that even if the three types of judges have, on average, very different standards, this fact has no bearing on measurement precision, and a similar statement holds for the two types of items. Although seemingly paradoxical, this does make sense if each replication involves the same three strata and the same two item types. Under these circumstances, any difference among means over replications is not attributable to judges' strata and/or item types because these two facets are fixed over replications.

This does not mean, however, that variance attributable solely to strata, item types, or their interaction is necessarily of no consequence to an investigator. For example, a large value for $\sigma^2(s)$ suggests that different types of judges tend to use different types of standards. This may not be a reliability issue in the sense that it affects *SE*, but it could be a validity issue for an investigator who believes that different types of judges should have the same standard, especially if the investigator has a theory to support that belief. This is an example of how generalizability theory blurs distinctions between reliability and validity.

Judges

In the context of standard setting, no facet has received more attention than judges. Indeed, some discussions make agreement among judges the sole criterion for an adequate standard. Such overemphasis is not simply unfortunate; it is grossly misleading. However, the judge facet is undeniably important in evaluating the adequacy of a standard-setting process.

Population/Sample Issues

To characterize the judge facet, it is necessary to specify clearly the population of potential judges, which may or may not be stratified. Specifying the population is almost always a political question for which there is no psychometrically correct answer. In particular, there is nothing wrong psychometrically with a population that consists solely or partially of stakeholders who are not experts in the subject matter assessed. However, the interpretability of the outcomes of the process clearly depends on the nature of the population. It is usually not too surprising that different populations of judges produce different outcomes [i.e., $\sigma^2(s) > 0$].

Once a population is specified, a plan for sampling from it needs to be developed and implemented. As complex as this may be, the theory and practice of sampling are generally well understood. Of course, if a convenience sample is employed, it is difficult to make clear arguments about generalization to a well-specified population.

Interjudge Issues

A central aspect of any traditional or generalizability analysis of the results of a standard-setting study focuses, directly or indirectly, on agreement among judges.[5] If judges within strata tend to have the same standards, $\sigma^2(j:s)$ will be small and SE will be smaller than it otherwise would be. If, in addition, $\sigma^2(s)$ is small, the investigator can confidently generalize over different types of judges--that is, the investigator may confidently state that the outcomes are invariant over different types of judges.

The preeminent consideration given to interjudge consistency is in itself a strong motivation for devising strategies to increase consistency among judges. However, artificial strategies for doing so need to be avoided. In particular, procedures that force consensus may reduce random error (SE) but increase systematic error that reduces validity, especially if the "consensus" is primarily a reflection of one dominant judge.

Often agreement among judges (or lack thereof) is indexed and reported as an interrater reliability (IR) coefficient. One expression for this coefficient, in terms of variance components from the hypothetical study, is[6]

$$IR = \sigma^2(i:t) / \left[\sigma^2(i:t) + \sigma^2(ji:st)\right]. \qquad (2)$$

Clearly, IR increases as $\sigma^2(ji:st)$ decreases and/or $\sigma^2(i:t)$ increases. However, note that both of these variance components are incorporated in SE in Equation 1. So even a large value for IR (suggesting considerable agreement among judges) does not guarantee a low value for SE. Indeed, the magnitude of IR has no necessary relationship with the magnitude of SE. For example, if $\sigma^2(i:t) = \sigma^2(ji:st)$, then $IR = .5$ no matter how large or small $\sigma^2(i:t)$ and $\sigma^2(ji:st)$ may be, whereas SE is directly affected by the magnitude of these two variance components.

Investigators and policymakers may have good reasons for instituting procedures to help ensure large values for interrater reliability coefficients. However, estimates of interrater reliability should not be offered as substitutes for, or indicators of, estimates of standard errors for cut scores. (Brennan & Lockwood, 1980, and Kane & Wilson, 1984, consider some other interrater reliability issues for standard setting.)

Items

The item facet contributes to error, and items are characterized as a random facet, whenever an investigator wants to generalize from the sample of items actually rated by judges to a larger (presumably infinite) universe of potential items. This seems to be a sensible assumption even if all items in a currently available pool are rated by the judges. The current pool is almost always but a sample of a very much larger pool that could exist.

---

[5] The statistics employed vary considerably. The traditional percent of agreement (with or without a chance correction) may be used. For a generalizability analysis, variance components reflecting judges' consistency would be reported. Detailed consideration of differences among these types of statistics is beyond the intended level of detail of this section.

[6] Technically, IR is an intraclass correlation. In this case, for items of a particular type, IR in Equation 2 is approximately equal to the product-moment correlation of the item ratings of two randomly selected judges from the same stratum.

As Equation 1 indicates, variability among items and the interactions of items with other facets contribute to *SE*. The variance component $\sigma^2(i{:}t)$ is the average (over item types) of the mean ratings of items of a particular type. As such, $\sigma^2(i{:}t)$ is analogous to the variance of item difficulties.

The variance component $\sigma^2(si{:}t)$ is the average, over item types, of the interactions of strata and items. This variance component will be large if different types of judges rank order items differently.

As noted previously, the variance of the item-type means, $\sigma^2(t)$, does not contribute to *SE* because all item types would be present in all replications. Stated differently, item types are fixed in that there is no generalization beyond the $n_t = 2$ item types actually employed. However, the interaction of item types with judges, $\sigma^2(tj{:}s)$, does contribute to *SE*. This variance component is large if, for at least some strata, the rank order of the judges' mean ratings (over items) differs by item type.

Although $\sigma^2(t)$ does not contribute to *SE*, a relatively large value for $\sigma^2(t)$ is sometimes viewed as a source of systematic error and, therefore, a threat to validity. Suppose, for example, that the two types of items are multiple choice (MC) and extended response (ER), and that judges assign a higher average rating to ER items than to MC items.[7] Judges then have a different standard for ER and MC items. This can be a threat to validity if an investigator or policymaker takes the position that MC and ER items should result in the same standard. Matters become still more complicated when examinee performance is taken into account, because then there can be various degrees of congruity/incongruity among judges' ratings, investigator judgments, and examinee performance on the two item types.

Occasions

For the results of any standard-setting process to be usable for practical purposes, the outcomes must be viewed as generalizable over occasions similar to the one when the judges ratings were obtained. If standard-setting outcomes were known to be idiosyncratic to the particular time when the ratings were obtained, they would not be given much credence.

Somewhat paradoxically, the occasion facet is probably the most obvious source of measurement error and the most difficult to treat in the sense of obtaining unconfounded estimates of its effects. Strictly speaking, direct estimation of intrarater reliability,[8] as well as variance components attributable to occasions and their interactions with other facets, requires replicating the standard-setting process with the same judges on at least two different occasions.[9] This is usually very difficult and, therefore, is seldom done.

---

[7] The obvious complexities of rating ER items are beyond the intended scope of this section.

[8] In this section, intrarater reliability refers to the correlation of a judge's ratings across occasions. In other contexts, the phrase "intrarater reliability" sometimes refers to a measure of consistency in item ratings for a particular judge on a specific occasion.

[9] The magnitude of intrarater reliability bears no necessary relationship with the magnitude of a standard error for cut scores. This follows from the fact that *all* variance components in *both* the numerator and the denominator of an intrarater reliability coefficient contribute to the standard error in the cut score(s) determined by a standard-setting process--assuming, of course, that items, judges, and occasions are all random facets.

In some standard-setting processes, judges go through two or more rounds of ratings, and only the final round's rating is used to compute the cut score. In such cases, ratings from the final two rounds can be used to obtain estimates of intrarater reliability, as well as estimates of variance components associated with occasions; but these estimates are likely to be somewhat biased by the judges' experience between rounds. Still, estimating intrarater reliability and variance components associated with occasions is usually one of the few reasonable uses of prior-rounds data in assessing error in a standard-setting outcome.

Minireplications

The occasion facet and the notion of replications are closely linked, but they are not isomorphic. What might be called "true" replications involve different conditions of several facets, including occasions. Some estimation problems are much more amenable to straightforward solutions when true replications are available. In particular, the standard error of the mean can be estimated directly without the statistical assumptions (e.g., uncorrelated effects) necessary to derive *SE* in Equation 1. For example, in an actual standard-setting process, judges may not act in a strictly independent manner. Thus, there can be flaws in the usual procedures for estimating the variance components for *SE* in Equation 1. When actual replications are available, these problems do not arise in estimating the standard error for the cut scores.

To get some of the statistical benefits of replications without the price of true replications, "minireplications" can sometimes be used. Minireplications are defined here as replications not involving the occasions facet or, more correctly, replications for a fixed occasion. The simplest example is two half samples, which involves splitting the item pool in half and having each half evaluated by a random half of the judges. The variance of the means for these two half samples can be used to estimate the standard error of the mean, when items and judges are viewed as facets in the universe. Such an approach is analogous to split-half procedures in classical test theory.

An obvious disadvantage of minireplications is that in most practical circumstances there can be only a very small number of them--usually only two. However, the advantage already discussed, as well as another one considered in another section, may be compelling enough to justify using minireplications.

An Expanded Conceptualization of Facets and Universes

Distinctions have been drawn between conditions of random facets (judges, items, and occasions) in a universe and fixed conditions (strata for judges and item types) present in each replication. These fixed conditions themselves can be called facets. In principle, different conditions of random facets are sampled from the universe for each replication. In contrast, all conditions of each of the fixed facets are present for each replication. That is, strata and item types are standardized conditions of measurement. Using terminology originally introduced by Kane (1982), the "universe of allowable observations" consists of both the random facets and the fixed, standardized conditions common to each replication. (This is the universe most closely associated with *SE* in Equation 1, although *SE* does not specifically incorporate variability over occasions.)

Kane (1982) uses the term "universe of generalization" for a universe that involves issues of generalization over both the random facets and the conditions that are fixed in the universe of allowable observations. So if an investigator wants to consider the consequences of generalizing over types of judges (strata) and/or item types, the investigator is referencing the universe of generalization in Kane's terminology. In this sense, the universe of generalization is associated with validity-related questions, whereas the universe of allowable observations[10] is associated with reliability-related questions. Again, generalizability theory blurs distinctions between reliability and validity.

Another facet that might be considered is the set of possible standard-setting methods. For the universe of allowable observations, it seems sensible that only one (fixed) method be present. For Kane's universe of generalization, however, method might be a facet, with generalization considered over all types of methods. Statements about such generalizations might involve divergent/convergent validity arguments.[11]

I have drawn distinctions between a universe of allowable observations and a universe of generalization to differentiate between certain reliability and validity issues. I am not arguing that it is sensible for different types of judges (strata), different item types, and/or different methods to yield the same standards. I am claiming, however, that if an investigator wants to make arguments about standard-setting outcomes for different types of judges, items, and methods, then such arguments are properly viewed in the context of validity, not reliability and random errors of measurement, given the conceptual framework proposed here.

To say that such arguments are "in the context of validity" does not mean that a particular standard-setting procedure is invalid if the outcomes are different for different types of judges and/or items. Such a result merely suggests that an investigator cannot dependably generalize from results for one item type to another and/or one type of judge to another. This would be evidence supporting an "invalidity" argument only if there were some accepted theoretical rationale supporting a claim that outcomes should be the same for different types of items and/or judges.

Clearly, if standard-setting outcomes are comparable over different types of judges, items, and/or methods, an investigator can make broader generalizations. It is a mistake, however, to associate the word "valid" with broader generalizations and the word "invalid" with narrower generalizations, unless there is a compelling theoretical argument that supports doing so. In the end, the crucial issue is that public pronouncements about standard-setting outcomes should be stated in a manner consistent with the degree of generalizability warranted by the data and any associated supportable theoretical arguments.

OTHER ISSUES

The previous two sections have laid out a framework for viewing standard setting from the perspective of generalizability theory. Because this is only a framework, many issues have not been treated in detail. In addition, no single perspective can capture all of the relevant issues. Therefore, a few additional issues follow that seem especially important or misunderstood.

---

[10] Cronbach et al. (1972) and Brennan (1992a) do not draw this distinction explicitly. They use the term "universe of generalization" in a way that seems to encompass both of Kane's (1982) possible universes. That is why, to this point, only the word "universe" (unmodified) has been used. A full discussion of universes would incorporate the Cronbach et al. "universe of admissible observations."

[11] Beliefs about different methods yielding common outcomes are probably misguided or overly optimistic (see Jaeger, 1989). Even in the fairly well-developed field of psychophysics, different methods of obtaining human judgments frequently lead to different results--even different "laws" (see Falmagne, 1985, pp. 315ff).

Whose Standards?

Because judges are central to a standard-setting process, it is understandable that there is a tendency to refer to the judges as "the standard setters." In almost all cases, however, judges' ratings (or functions of them) are more properly viewed as advice to decision/policymakers who, by virtue of their positions, are at liberty to take, modify, or neglect that advice as they see fit. If the judges' advice is taken, then the conceptual framework and notions of error introduced previously are sensible. They are less sensible if the judges' advice is modified and probably nonsensical if the judges' advice is neglected.

Definitions

Currently, definitions are most clearly associated with NAEP achievement-level-setting processes. However, there are at least implicit definitions associated with even a single pass/fail cut point--namely, what it means to say an examinee passes or fails. Kane (1994) discusses this and related issues in his consideration of performance standards.

When the opportunity to modify definitions (even in minimal ways) is part of a standard-setting process, two complex issues are encountered. First, how can an investigator estimate or report what definitions would result if the process were replicated? No statistical approach seems particularly viable, except perhaps the use of minireplications. Second, how does an investigator address (let alone ensure) the congruence of definitions, cut scores, and exemplars? The most obvious answer is to assume that the judges are responsible for ensuring congruence because the judges are involved in generating all the outcomes. This may be minimally acceptable, but it is not likely to be terribly compelling.

Variability in definitions is probably the least examined aspect of standard setting because it is so difficult to study this issue. However, the interpretations associated with cut scores are intimately associated with the definitions. Consequently, confidence in such interpretations is delimited by any ignorance about variability in definitions.

One route out of this dilemma, of course, is to specify the definitions a priori so that they are not subject to change by judges; but then the judges must be willing to adopt the definitions as their own.

Exemplars

The process of choosing exemplars usually involves some consideration of examinee performance on the items. For example, the first step in identifying exemplars may involve a statistical criterion such as including an item as a possible exemplar only if the probability is at least X that examinees at every score level in a range of scores get the item correct. Then judges are asked to choose exemplars from the set of items that satisfy this criterion.

Although the criterion is often more complicated, even in this simplified example it is evident that (a) choosing X is a decision in the purview of policymakers, not psychometricians; and (b) the choice of X can have a substantial impact on the final exemplars chosen, especially if the pool of potential exemplars is limited, as it often is.

Clearly, then, the rationale for choosing X should be stated as openly as the rationale for other a priori decisions about the standard-setting process (e.g., choice of types of judges, types of items, or method). Further, efforts should be made to inform users about what it means for an item to be an exemplar and what X means relative to the exemplars. For example, if an item is used as an exemplar for, say, "advanced," and X is .65, users should be disavowed of conclusions such as, "All advanced examinees get this item correct, because the item is exemplary," and "Advanced examinees get 65% of the items correct."

Weighting

In a standard-setting study that involves different types of items and/or judges, item types and/or judges' strata of necessity are weighted (perhaps only implicitly) in arriving at a standard. For example, if there are two types of items in a pool and the standard is based on the simple average of the ratings over items, then the weights for the two types of items are effectively the proportions of the two types of items in the pool. Alternatively, explicit, a priori weights could be assigned to the item types--weights that need not equal the proportions of items in the pool--in which case, formulas for standard errors would become more complex than, for instance, Equation 1.

Weighting can also affect a standard if different degrees of importance are attached to content categories in a framework. For example, content experts might decide a priori that content categories A, B, and C should be weighted .5, .3, and .2 in arriving at examinee scores. In that case, unless there was a compelling argument to the contrary, it would seem sensible that judges' ratings of items from these three categories should receive comparable weights.

Weighting almost always affects to some extent the results of a standard-setting process, and the influence of weighting is often unknown to many who use the resulting standards. Further, weighting issues are sometimes obscure even to those who develop the standards. In principle, this may be problematic if the weights have a substantial effect. From a pragmatic perspective, however, an investigator may sometimes decide that introducing weighting issues into a standard-setting study will complicate matters too much.

Unit of Analysis

Because the tests on which standards are set consist of items or tasks, it seems natural to develop and discuss models (mathematical or verbal) at the level of items or tasks. However, it is not necessarily sensible to rely heavily on item-level or task-level analyses in evaluating the characteristics of the resulting standards. For example, even if judges are inconsistent in their evaluations of particular items, the same judges may be quite consistent when their evaluations are averaged over the set of items or tasks on which the standard is based. Aggregated evaluations or judgments are the principal concern; item-level or task-level analyses can be misleading, especially if they are overemphasized.

Dimensionality

Issues of dimensionality often underlie discussions of various aspects of standard-setting. This occurs in part because unidimensional models are sometimes used in a standard-setting process when various arguments can be put forward suggesting that the unidimensionality assumption is violated. Even if this assumption is violated, however, the standard-setting results may be defensible if they can be shown to be robust with respect to violations of the unidimensionality assumption.

A more subtle, but nonetheless important, dimensionality consideration is the rather pervasive, implicit assumption that the dimensionality of examinee item responses is, in some sense, "truth" that content experts and judges' ratings should reflect. This assumption seems unwarranted.

At least three potentially different dimensionalities may be involved in developing, using, or interpreting standard-setting results. First, those responsible for establishing the content framework typically use a set of dimensions that they believe to be useful for characterizing the content area and for developing test specifications. Second, correlational or covariance matrices of examinee responses can be subjected to various statistical procedures that result in characterizing the dimensionality of examinee responses. Third, in principle, the same types of statistical procedures can be applied to judges' ratings to characterize the dimensionality of such ratings.

These three "dimensionalities" need not correspond. For example, in the absence of any compelling argument to the contrary, there is no reason to believe that examinee response dimensions should mirror the content framework dimensions. More important for the purposes of this paper, however, there is no reason to assume a priori that the dimensions associated with examinees' responses and the dimensions associated with judges' ratings should or will be isomorphic (see Kane, 1987). It is entirely possible that what judges think examinees should be able to do reflects a conceptualization of the interrelationships among content areas that is not evident in what current examinees can do.

Scale Scores

In the initial discussion of the outcomes of a standard-setting process, it was noted that an average of judges' ratings is usually not the cut score *per se*. Rather, the average of the ratings is transformed to a scale score, and this particular scale score becomes the cut score (assuming policymakers do not modify it). In a sense, then, standard errors such as those in Equation 1 are not grounded in the most appropriate metric (i.e., the scale scores used to make decisions about examinees), and standard errors on the scale score metric are usually necessary for users to evaluate standards. Obtaining standard errors on the scale score metric is often a statistically challenging task, especially if the score scale is a nonlinear transformation of raw scores. The process is simplified considerably if actual replications, or minireplications of the type discussed previously, are available. For example, if half samples are employed, the mean for each sample can be transformed to its corresponding scale score, and from these scale scores the standard error of the scale score mean can be computed directly.

Because standard setting and scaling are typically viewed as different psychometric issues, there is a natural tendency to assume that one does not affect the other. Usually, it is true that standard setting does not affect scaling, although a persuasive argument can be made that a score scale should be defined in a manner that accommodates any standard setting that may be undertaken. By contrast, scaling almost always influences standard setting in the sense that one of the results of a standard-setting activity is a cut score on the score scale. Clearly, then, the interpretability and meaningfulness of the cut score are intimately tied to the characteristics of the score scale. This can be particularly problematic if cut scores are at the extremes of a score scale where empirical evidence about scale score characteristics may be sparse.

The interplay between standard setting and scaling is also evident in statistics such as the percentage of examinees above a cut score. Suppose this percent increases noticeably over two years. Clearly, this could be a consequence of (a) "true" improvement in examinee performance; and/or (b) measurement error in examinee scores in the region of the cut score; and/or (c) equating error. The increase could also be viewed as attributable partly to error in establishing the standard, because a somewhat different specification of the standard (still within, say, a standard error band) might result in a different percentage increase.

A standard-setting process, no matter how well it is implemented, inevitably places an additional burden on a score scale. Also, the interpretability and usefulness of standard-setting outcomes depend, at least in part, on the score scale.

SOME SUGGESTED STANDARDS

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1985) contains several *standards* that relate to cut scores for making decisions about individuals (e.g., 1.24, 2.10, 6.9, and 11.3), but otherwise *standards* for standard setting are considered only minimally. In this section, a few *standards* for standard setting are suggested. These *standards* have two notable characteristics.

First, these *standards* are purposely stated at a level of detail that makes them applicable to virtually any standard-setting process. *Standards* for specific processes seem inadvisable. Conditioning *standards* on a single method runs the serious risk of deifying that method inappropriately or prematurely. More important, the basic principles of generalizability discussed in this paper are sufficiently broad to be applicable to multiple methods. Of course, special issues arise with particular methods, and these issues should be given the attention they deserve in technical documentation. However, *standards* need to be pitched at an appropriate level of specificity. Too little specificity renders *standards* useless; too much turns them into a cookbook that inhibits innovative improvements. In short, for the purposes of specifying *standards* it is neither necessary nor desirable to restrict attention to a single method or class of methods.

Second, not one of these *standards* makes reference to any desirable numerical value for a statistic. There is a long history in measurement and statistics of specific numerical values serving as benchmarks for judging adequacy (e.g., the widespread use of .80 as "adequate" reliability and of .05 as an "appropriate" level of statistical significance). However, there is very little, if any, compelling rationale that supports these benchmarks. Judgments of statistical adequacy are just that--judgments. The goal should be that these judgments be well informed and properly influenced by the context and use of the statistics, not that they be made with reference to arbitrary benchmarks. Of course, such a position can be uncomfortable for policymakers. Grappling with the reality of the process, its outcomes, and its consequences may be difficult, but it is much more likely to lead to sound decisions than using questionable statistical benchmarks.

The following suggested *standards* are grouped together, but they are not listed in order of importance.

1. The characteristics of the judges, the items, and the standard-setting process should be described at an appropriate level of detail for the various types of users.

   Often there are several types of users, which means that descriptions are needed at different levels of detail. At least one level should be sufficiently detailed so that an independent investigator could replicate the process.

2. Clear distinctions should be drawn between the role of policymakers and the role of judges who provide ratings.

   For example, a priori conditions specified by policymakers (e.g., types of judges to be employed, method to be used, types of items evaluated, extent to which judges are constrained in any way) should be reported.

   Also, if the final standard is any modification of the judges' ratings (except a scale transformation), the nature of such a modification and the reasons for it should be reported.

3. For each of the outcomes of a standard-setting activity, information should be provided that addresses the stability, replicability, or trustworthiness of results.

   Standard-setting activities generally have one or more of the following outcomes: definitions, cut scores, and exemplars. Current literature emphasizes stability of results primarily for cut scores. However, because definitions and exemplars can be a crucial component of interpretations, it is important that some type of evidence be provided to support the trustworthiness of the definitions and exemplars.

4. Primary results should be reported at the appropriate level of aggregation.

   For example, for the achievement-level-setting procedures in NAEP, standards are set using the mean score over judges and items. Standard errors, therefore, should be reported for such means. Standard errors or other statistics for item- or task-level analyses may be informative as secondary information, but they are not definitive.

5. For any estimated standard errors, the facets that are random and the facets that are fixed should be specified explicitly.

   All other things being equal, the larger (smaller) the number of random facets, the larger (smaller) the estimated standard error. Consequently, a standard error cannot be interpreted clearly without an understanding of which facets are random and which are fixed.

6. Important statistics, especially standard errors, should be reported for the metric used to make decisions.

   Standard errors are most easily computed for a raw score metric, and it can be difficult to estimate standard errors for the scale score metric, especially if it is a nonlinear transformation of raw scores. Even so, standard errors for the scale score metric should be reported. If doing so requires some simplifying assumptions, they should be specified.

7. If interrater and/or intrarater reliabilities are reported, their importance should not be exaggerated, and they should not be offered as substitutes for standard errors.

   High interrater and/or intrarater reliability does not guarantee low standard errors of measurement for cut scores that are means of judges' ratings over items.

8. Any unanticipated aspects of the standard-setting process, or anomalous outcomes, that might affect interpretations should be reported.

   For example, a serious disruption at the site of the standard-setting process might be reasonable cause for reduced confidence in results. As another example, consider the NAEP achievement levels for the 4th, 8th, and 12th grades. Clearly, if the 8th-grade advanced cut score was higher than the 12th-grade advanced cut score, users might reasonably question the trustworthiness of the cut scores.

9. Users should be cautioned about any reasonably anticipated misuses of the outcomes of a standard-setting process.

   Standard setting is a difficult activity, involving many a priori decisions and many assumptions. Consequently, the outcomes can be misunderstood. It is particularly important that users be disavowed of any belief that standard-setting outcomes are anyone's "truth." In addition, if exemplars are an outcome, great care should be taken in informing users about how exemplars should (and should not) be interpreted. This standard *has the spirit of one of the codes (Test Developer Code 11) of the Joint Committee on Testing Practices (1988).*

CONCLUDING COMMENTS

Standards are not "truths," but they can be more or less defensible and useful depending on how carefully they are established, the extent to which the outcomes are generalizable, and the degree to which interpretations are clear and well supported. Clear documentation is crucial to a defense. This documentation needs to delineate, among other things, the respective roles of judges and policymakers, and it needs to specify which facets are random and which are fixed in computing any estimates of error.

From the perspective of generalizability theory, it is important that, directly or indirectly, estimates of variability over replications be reported for cut scores (as well as for definitions and exemplars, to the extent possible). It should not be assumed that estimates of interrater and intrarater reliability serve this purpose. Indeed, the magnitudes of interrater and intrarater reliabilities have no necessary relationship with standard errors of cut scores.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Brennan, R. L. (1992a). *Elements of generalizability theory* (Rev. ed.). Iowa City, IA: American College Testing.

Brennan, R. L. (1992b). Generalizability theory. *Educational Measurement: Issues and Practice, 11,*(4), 27-34.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement, 4,* 219-240.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Falmagne, J. C. (1985). *Elements of psychophysical theory*. New York: Oxford University Press.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: American Council on Education/Macmillan.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). New York: American Council on Education/Macmillan.

Joint Committee on Testing Practices. (1985). *Code of fair testing practices in education*. Washington, DC.

Kane, M. (1982). A sampling model for validity. *Applied Psychological Measurement, 6,* 125-160.

Kane, M. (1987). On the use of IRT models with judgmental standard-setting procedures. *Journal of Educational Measurement, 24,* 333-345.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64,* 425-461.

Kane, M., & Wilson, J. (1984). Errors of measurement and standard setting in mastery testing. *Applied Psychological Measurement, 8,* 107-115.

Livingston, S. A. (1995). Standards for reporting the educational achievement of groups. *In Proceedings of Joint Conference on Standard Setting for Large-Scale Assessments*. Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Standard Setting From the Perspective of Generalizability Theory

Summary of Break-out Session[1]

The discussion focused first on the *replication of the standard-setting process*. The question that guided the discussion was. Should different methods be used when replicating the standard-setting process? Brennan recommended the use of the same method each time. He suggested that different methods yield different results and, therefore, would generate a large index of error over replications. These threaten the validity of the process.

A participant posed another question: Might different steps in an iterative standard-setting process be viewed as separate "occasions" (in the sense used in Brennan's paper)? Should there be high variability between iterations? Brennan replied that one should view the entire process as one facet, because the final result depends on the final stage only. He pointed out that there should be high intra-rater variability between rounds and between the final stage and the stage immediately preceding it, if the process is to be a good indicator of the reliability of the process.

He indicated that it is flawed reasoning to assume that high inter-rater reliability is a necessary condition, and that if inter-rater reliability is high, the standard error will be low. He emphasized that in standard setting, the final product is averaged over the judges, and that in the formula for the standard error, there are five standard error terms accounting for a multiplicity of judges being selected from various strata in the population and the presence of several item types. He pointed out that an inter-rater reliability coefficient takes into account two of the five terms--the average of the variances of the items of a particular type and the average of the judge-item interactions within each stratum-type combination. He emphasized that *the standard error is a more important statistic than inter-rater reliability, although that should not be ignored.*

It was pointed out that it is hard not to confound judges and occasions as facets because *different judges are often used for replications of the standard-setting process*. If there are different replications using different judges, one does not have to deal with variance components; one can calculate the standard error of the means of the replications. This process eliminates having to deal with not satisfying statistical assumptions, because the replications are independent.

Brennan was asked how he reconciled his position related to de-emphasizing inter-rater reliability. Brennan prefaced his response by saying that the training of judges is very important. He added that weighting individual judges' contributions differently is politically impossible, once the judges have been included in the process. He further explained that having the judges work from a common conceptual framework is important, because it will reduce variability between judges; but he emphasized that some methods will result in weighting judges differentially. He posed an example: A panel may be composed of 55 percent teachers, 30 percent nonteacher educators, and 15 percent from the general public. The implicit weights may not be obvious to policymakers or even to technical persons involved in the process. The contribution of these unequally weighted types of judges, however, is not an issue in

---

[1] This is a summary of the break-out discussions for Robert Brennan's presentation. The session was facilitated by Sharif Shakrani (National Center for Education Statistics) and recorded by Arnold Goldstein (National Center for Education Statistics).

BEST COPY AVAILABLE

calculating the overall variance. He emphasized that *in replicating the standard-setting process, the issue is not whether different judges yield different results. The issue is whether use of a different process would yield different results. Replications do tend to give similar results in practice. It was pointed out, though, that there is no agreement on what constitutes a replication.*

Brennan addressed the practice of some judges inflating or lowering standards to support their perceptions that the standard being set is too high or too low. He indicated that this is a defensible practice and a policy judgment. He noted that there is no inherent virtue in panelists operating either as individuals or in concert; it is a policy judgment whether judges should have feedback that would influence them to change their original concepts. Brennan advised that *more research is needed to inform and advise policymakers about the necessary qualifications of judges.* He emphasized that standards should not be set for fourth grade, for example, by people unfamiliar with fourth-grade instructional content and developmental issues. He concluded by saying that determining the sampling frame for selecting judges is itself a policy judgment.

The discussion addressed another question on how to report variance and standard errors when both dichotomous and polytomous items are used on a test. How should the variance and standard error be derived? It was suggested that a single standard error should be reported that includes both types of items. One participant pointed out that according to the National Academy of Education, a standard-setting process that yields different standards for the two types of items is invalid.

The participants addressed another question: How reliable can a test be when only a portion of a domain is tested? It was suggested that 8 to 12 open-ended items would result in good coverage and adequate reliability. The group agreed that matrix sampling allows for the inclusion of a large number of items.

The group discussed briefly the application of the *opportunity-to-learn standards*. One questioner asked if state performance tests may be biased because students did not have the opportunity to cover the content? The group agreed that this was not bias; they explored the meaning of instructional validity.

The discussion prompted another question: Which method of standard setting offers the highest reliability? Brennan said that errors of measurement (standard error) are more relevant to consider than reliability when assessing standard-setting methods. He observed that the Angoff method yields the lowest number of standard errors of all the methods, but the chosen method should be picked to satisfy the policymakers' purpose; it should have a generally accepted research base and credibility among stakeholders, including the courts. He pointed out that *psychometrics does not indicate which is the best method in a given situation*.

One participant observed that *relatively little research has been done on examinee-centered methods.* It was suggested that, to date, research has been conducted mostly on item-judgment methods.

Standards-Based Score Interpretation: Establishing Valid Grounds for Valid Inferences

Samuel Messick[1]

Vice President for Research, Educational Testing Service, New Jersey

ABSTRACT

The construct validity of content standards is addressed in terms of their representative coverage of a construct domain and their alignment with the students' cognitive level of developing expertise in the subject matter. The construct validity of performance standards is addressed in terms of the extent to which they reflect increasing levels of construct complexity as opposed to construct-irrelevant difficulty. Also critical is the extent to which performance standards characterize the knowledge and skills operative at each level both to accredit specific accomplishment and to serve as goals for further learning. All of this depends on construct-valid assessment attuned to the content standards and the development of dependable scoring rubrics and measurement scales for representing the performance standards.

In standards-based education reform, a lot depends on the establishment of valid standards specifying both the critical content and the desired performance levels of student accomplishment, because these two kinds of standards compose the driving force that energizes the reform movement. Basically, content standards specify what students should know and be able to do; performance standards specify the level and quality of that knowledge and skill that are deemed acceptable. Appraising whether assessed student competence meets a performance standard requires that the two be compared as points, as it were, on the same measurement scale. As a consequence, the validity of these standards cannot be separated from the validity of the assessment itself. That is, a construct-valid measurement scale of some sort is needed because without it there is not only no assessed content competence, but there is also no performance standard.

Hence, to address the construct validity of both content standards and performance standards, the same framework of validity criteria and forms of evidence must be used as are needed to appraise the construct validity of assessed student competence. This is necessary because the construct validity of the content standards and the construct validity of the measurement scales go hand in hand. Moreover, to be meaningfully interpreted and reported, both the assessed competence and the performance standard must be described in the same *construct* terms. That is, a performance standard has two critical aspects: its location on the measurement scale and its meaning in terms of the nature or quality of the knowledge and skill characterizing proficiency at that level. As a consequence, the construct validity of the meaning of the performance standard as well as that of the assessed competence, both being interpreted points on the same measurement scale, must be evaluated in the same *evidential* terms. Because the location of the performance standard is fundamentally a matter of value judgment, its validity must be addressed in terms of the reasonableness of the procedures used for determining it.

---

[1] I gratefully acknowledge helpful comments on the manuscript provided by Ann Jungeblut, Robert Mislevy, and Michael Zieky.

Next, the criteria or standards of validity are briefly reviewed, as are the forms of evidence pertinent to the construct validation of any assessment, including performance assessment. Then, these general validity principles are applied to a consideration of the construct validity of both content standards and performance standards, drawing implications as well for evaluating the standard-setting process whereby performance standards are determined.

STANDARDS OF VALIDITY

Broadly speaking, validity is nothing less than an evaluative summary of both the evidence for and the actual, as well as potential, consequences of score interpretation and use (i.e., construct validity conceived comprehensively). This comprehensive view of validity integrates considerations of content, criteria, and consequences into a construct framework for empirically testing rational hypotheses about score meaning and utility. Fundamentally, then, score validation is empirical evaluation of the meaning and consequences of measurement. As such, validation combines scientific inquiry with rational argument to justify (or nullify) score interpretation and use. Hence, validity becomes a unified concept that integrates multiple supplementary forms of convergent and discriminant evidence.

However, to speak of validity as a unified concept does not imply that validity cannot be usefully differentiated into distinct aspects to underscore issues and nuances that might otherwise be downplayed or overlooked, such as the social consequences of performance assessments or the role of score meaning in applied test use. The intent of these distinctions is to provide a means of addressing functional aspects of validity that help disentangle some of the complexities inherent in appraising the appropriateness, meaningfulness, and usefulness of score inferences.

Aspects of Construct Validity

In particular, six distinguishable aspects of construct validity are highlighted here as a means of addressing central issues implicit in the notion of validity as a unified concept. These are content, substantive, structural, generalizability, external, and consequential aspects of construct validity. In effect, these six aspects, briefly characterized as follows, function as general validity criteria or standards for all educational and psychological measurement (Messick, 1989, 1994b):

1. The content aspect of construct validity includes evidence of content relevance, representativeness, and technical quality (Lennon, 1956; Messick, 1989).

2. The substantive aspect refers to theoretical rationales for the observed consistencies in test responses, including process models of task performance (Embretson, 1983), along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks.

3. The structural aspect appraises the fidelity of the scoring structure to the structure of the construct domain at issue (Loevinger, 1957).

4. The generalizability aspect examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks (Cook & Campbell, 1979; Shulman, 1970), including validity generalization of test-criterion relationships (Hunter, Schmidt, & Jackson, 1982).

5. The external aspect includes convergent and discriminant evidence from multitrait-multimethod comparisons (Campbell & Fiske, 1959), as well as evidence of criterion relevance and applied utility (Cronbach & Gleser, 1965).

6. The consequential aspect appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice (Messick, 1980, 1989).

In one way or another, these six aspects seek evidence and arguments to discount the two major threats to construct validity--namely, construct underrepresentation and construct-irrelevant variance--as well as to evaluate the action implications of score meaning. In construct underrepresentation, the assessment is too narrow and fails to include important dimensions or facets of the construct. In construct-irrelevant variance, the assessment is too broad, containing excess reliable variance that is irrelevant to the interpreted construct. Both threats are operative in all assessment. Hence, a primary validation concern is the extent to which the same assessment might underrepresent the focal construct while simultaneously contaminating the scores with construct-irrelevant variance.

Validity as Integrative Summary

The six aspects of construct validity apply to all educational and psychological measurement, including performance assessments. Taken together, they provide a way of addressing the multiple and interrelated validity questions that need to be answered in justifying score interpretation and use. As I have maintained in previous writings, it is "the relation between the evidence and the inferences drawn that should determine the validation focus" (Messick, 1989, p. 16). This relation is embodied in theoretical rationales or in persuasive arguments that the obtained evidence both supports the preferred inferences and undercuts plausible rival inferences. From this perspective, as Cronbach (1988) concluded, validation is evaluation argument. That is, as stipulated earlier, validation is empirical evaluation of the meaning and consequences of measurement. The term "empirical evaluation" is meant to convey that the validation process is scientific as well as rhetorical and requires both evidence and argument.

By focusing on the argument or rationale employed to support the assumptions and inferences invoked in the score-based interpretations and actions of a particular test use, it is possible to prioritize the forms of validity evidence needed in terms of the important points in the argument that require justification or support (Kane, 1992; Shepard, 1993). Helpful as this may be, problems still remain in setting priorities for needed evidence because the argument may be incomplete or off target, because all of the assumptions may not be addressed, and because the need to discount alternative arguments evokes multiple priorities. This is one reason Cronbach (1989) stressed cross-argument criteria for assigning priority to a line of inquiry, criteria such as the degree of prior uncertainty, information yield, cost, and leverage in achieving consensus.

The point here is that the six aspects of construct validity afford a means of checking that the theoretical rationale or persuasive argument linking the evidence to the inferences drawn touches the important bases and, if not, an argument should be provided that such omissions are defensible. These six aspects are highlighted because most score-based interpretations and action inferences, as well as the elaborated rationales or arguments that attempt to legitimize them (Kane, 1992), either invoke these properties or assume them, explicitly or tacitly. That is, most score interpretations refer to relevant content and operative processes, presumed to be reflected in scores that concatenate responses in domain-appropriate ways and that are generalizable across a range of tasks, settings, and occasions. Further, score-based interpretations and actions are typically extrapolated beyond the test context on the basis of presumed or documented relationships with nontest behaviors and anticipated outcomes or consequences. The challenge in test validation is to link these inferences to convergent evidence that supports them as well as to discriminant evidence that discounts plausible rival inferences. Evidence pertinent to all of these aspects needs to be integrated into an overall validity judgment to sustain score inferences and their action implications, or else compelling reasons should be provided to justify omissions, which is what is meant by validity as a unified concept.

VALIDITY OF STANDARDS

Turning to the question of the validity of both content and performance standards, the construct validity of content standards is addressed in terms of their representative coverage of a construct domain and their alignment with students' cognitive levels of developing expertise in the subject matter. This treatment of the validity of content standards emphasizes the content, substantive, generalizability, and consequential aspects of construct validity. The construct validity of performance standards as interpreted score levels is addressed in terms of the extent to which they reflect increasing construct complexity as opposed to construct-irrelevant difficulty. This treatment emphasizes the structural, generalizability, external, and consequential aspects of construct validity.

Content Standards as Blueprints for Teaching and Testing

Content standards specify what students should know and be able to do in a subject area at their particular level of developing expertise in the field. Content standards are concerned with what students should know and be able to do in different years of study in a discipline, as in 4th-, 8th-, and 12th-grade mathematics or language arts. Within-grade levels of proficiency, as will be seen, are captured by performance standards.

Thus, content standards have a temporal dimension. They specify not only *what* knowledge and skills should be attained, but also *when*. Hence, the construct validation of content standards needs to address not only the relevance and representativeness of the *what* of the subject matter or construct domain, but also the appropriateness of the *when* to the students' cognitive levels of developing expertise. Judgments of what and when are usually combined by setting distinct content standards for different grade levels in a discipline, but these also need to be coordinated across grades to reflect an appropriate course of academic development. Ideally, both the substantive and the temporal aspects of content standards should be addressed in appraising their construct validity.

By specifying what students are expected to learn, content standards provide blueprints for what is important to teach as well as to test. However, content standards typically refer to generic constructs: For example, some content standards in eigth-grade mathematics require that students should "understand the process of gathering and organizing data"; further, they should "be able to calculate, evaluate, and communicate results" (Mullis, Dossey, Owen, & Phillips, 1993, p. 51). Learning exercises and assessment tasks are then selected or created to embody these generic processes. The validity of the content standards as blueprints for teaching and testing depends on the extent to which the learning exercises engender these processes and the assessment tasks tap them. Thus, the construct validity of the content standards and the construct validity of the assessment tasks are inseparable.

Although a separate topic not to be elaborated here, the construct validity of opportunity-to-learn standards and that of the learning exercises are similarly inseparable. The validity of opportunity-to-learn standards depends on exposure to learning experiences that are construct valid in the sense that they actually engender or facilitate the development of the knowledge and skills specified in the content standards.

The linking of content standards to assessment tasks shown to engage the specified processes bears on the substantive aspect of construct validity. A key issue of the content aspect is the extent to which the content standards circumscribe the boundaries and reflect the structure of the subject-matter construct domain. The major concern is to minimize construct underrepresentation. To be valid, the content standards and their associated assessment tasks should be both relevant to and representative of the construct domain. Hence, the content standards should specify (and the associated assessment tasks should sample) domain processes in terms of their functional importance.

A major problem is sorting out evidence of domain processes in complex tasks and especially disentangling focal construct processes from ancillary processes involved in task performance (Wiley, 1991). This problem is serious because ancillary processes, which are ordinarily numerous in complex task performance, are a potential source of construct-irrelevant variance.

Functionally important knowledge and skill in a subject-matter or construct domain may be addressed from at least two perspectives: what is actually done in the performance domain--for example, as revealed through techniques akin to job analysis--and what differentiates and characterizes developing expertise in the domain, which would usually emphasize different tasks and processes. The first perspective addresses the substantive aspect of content standards and the second addresses the temporal aspect.

In effect, the content standards specify the constructs that are not only to be taught but are also to be assessed in standards-based performance assessment. Further, as I have maintained elsewhere, "the meaning of the construct is tied to the range of tasks and situations that it generalizes and transfers to" (Messick, 1994a, p. 15). This leads to the generalizability aspect of construct validity and, in particular, to the distinction between generalizability as consistency or reliability and generalizability as transfer.

Brennan (1995) has discussed generalizability across judges, occasions, and tasks--topics typically subsumed under the heading of reliability--as well as generalizability across standard-setting methods. Also of concern are generalizability across measurement methods and scoring rubrics. Generalizability as reliability refers to the consistency of performance across the judges, occasions, and tasks of a particular assessment, which might be quite limited in scope. For example, there has been widespread concern that some assessments with a narrow set of tasks might attain higher reliability in the form of cross-task consistency at the expense of construct validity. In contrast, generalizability as transfer requires consistency of performance across tasks that are representative of the broader construct domain. In other words, transfer refers to the range of tasks that performance on the assessed tasks facilitates the learning of or, more generally, is predictive of (Ferguson, 1956). Thus, generalizability as transfer depends not only on generalizability theory but also on domain theory--that is, on the construct theory of the subject-matter domain. In essence, then, generalizability evidence is an aspect of construct validity because it establishes boundaries on the meaning of the construct scores.

Content standards are at the heart of standards-based education reform because they are presumed to have positive consequences for teaching and learning. Evidence documenting such positive outcomes bears on the consequential aspect of construct validity, which, for its full appraisal, also requires attention to the possibility of unintended adverse side effects. For example, establishing common content standards for all students is a selective process that privileges certain knowledge and skills over other possibilities. This might inadvertently lead, as Coffman (1993) indicates, to limitations on the development of those other skills and, hence, to unintended restrictions on the diversity of talent.

In effect, content standards for all students constitute a common denominator. To be sure, the impact of such a common denominator is most insidious when students are held to low standards, as in minimum-competency testing, rather than to high standards. Hence, the levels that students are challenged to reach become important as educational goals, which leads us directly to the topic of performance standards.

Performance Standards as Challenges or Hurdles

Performance standards refer to the level of competence students should attain in the knowledge and skills specified by the content standards as well as the form or forms of performance that are appropriate to be evaluated against the standards. To account for the differential complexity of information-processing requirements in different years of study in a discipline, which is attuned to the students' levels of developing expertise, standards of performance considered to be "good enough" are typically set separately by grade level. For example, in the National Assessment of Educational Progress (NAEP), performance standards for basic, proficient, and advanced levels are set separately for grades 4, 8, and 12.

By specifying the form or forms of performance that are appropriate to evaluate, performance standards essentially circumscribe the nature of the evidence relevant to deciding whether the standards have been met--for example, whether the evidence should be an essay, a project, a demonstration, a mathematical proof, a scientific experiment, or some combination of these. An important issue is whether standards-based score interpretation and reporting can legitimately be formulated in terms of the generic constructs of knowledge and skill specified in the content standards or whether they need to be specific to the method of measurement, that is, specific to knowledge and skill exhibited via a particular method.

If the latter specificity holds, interpretation is limited to construct-method units, which implies that there are not only distinct performance standards but also distinct content standards for each method of measurement. To attain the power of the former interpretation in terms of generic constructs requires evidence of generalizability across measurement methods.

This is a fundamental issue because the content standards evoke generic constructs of knowledge and skill that, if attained at the levels specified by the performance standards, are deemed relevant to a range of diverse problems and applications. That is, the content standards specify knowledge and skill that are considered important precisely because they are generalizable and transferable across problems and situational contexts, including measurement contexts. Hence, the degree of generalizability of the construct scores across measurement methods bears directly on the meaning of the constructs.

Moreover, attention should be paid not just to convergent evidence of consistency across methods but also to discriminant evidence of the distinctness of constructs within method. Such multiconstruct-multimethod comparisons are needed to help disentangle construct-relevant variance from construct-irrelevant method variance. But more about this later.

As ordinarily conceptualized, a performance standard is a point on a measurement scale, or a set of points on a profile of scales, or a region in a multidimensional space. A "softer" version of performance standards is associated not with cut points but with utility functions in a decision-theoretic approach to standard setting (van der Linden, 1995). For the sake of simplicity, a performance standard is characterized here only as a point on a scale because the argument about the centrality of the measurement scale in standard setting applies equally well to profiles and other multidimensional representations as well as to utility functions.

A measurement scale of some sort is critical to the setting and use of performance standards for at least two reasons. First, without a measurement scale, there can be no points on the scale and hence no performance standards. This is the case because the notion of performance standards implies an ordering of tasks (or of performances on a particular task) such that some of the performances are considered good enough and others not good enough. Such an ordering constitutes a rudimentary measurement scale. By taking into account the structure of interrelations among task or performance scores, more powerful, model-based measurement scales, such as the Item Response Theory (IRT)-based scales developed for NAEP, may be fitted to the data.

The second reason that not just a measurement scale but an interpreted-measurement scale is critical is that meeting a performance standard should not simply attest that the assessed performance is good enough. To be educationally useful, performance standards should also characterize the nature of the knowledge and skill entailed at that level as well as point to what needs to be accomplished for further mastery. One implication of this is that the measurement scale should extend beyond the level of the performance standard. Another implication is that various levels on the scale, especially the performance-standard levels, should be tied to process descriptions of what constitutes proficiency at each level.

The development of these process descriptions would be facilitated if the various levels were benchmarked by tasks for which students scoring at each level had a high probability of success while students at lower levels had less likelihood of performing well. This benchmarking is important because the interpretation and reporting of scores relative to performance standards require evidence, first, that tasks at a given scale level actually engage the knowledge and skill attributed to proficiency at this level and, second, that the performance of students at this level is validly characterized by the process description. Thus, the construct validity of the performance standards and the construct validity of the measurement scale are inseparable.

The construct validity of the performance standards as well as of the measurement scale, is vulnerable to threats of both construct underrepresentation and construct-irrelevant variance. For example, if acknowledged masters failed to meet the standard, construct irrelevancy in the measure would be suspected. Alternatively, if acknowledged nonmasters met the standard, construct underrepresentation would be suspected--the nonmasters might be proficient in the assessed part of a sparsely covered domain but less proficient in the unassessed part. The latter situation is the bane of selection testing and of criterion prediction more generally, as some individuals do well on the domain processes covered in the predictor tests but perform poorly on unmeasured processes important in criterion performance. This problem of construct underrepresentation is critical in standards-based educational assessment. Even NAEP, with balanced-incomplete block (BIB) spiraling, has trouble covering the important bases.

The identification of benchmark tasks and process descriptions is facilitated by development of the more powerful model-based measurement scales, to be sure, but much of performance assessment is limited to rudimentary scales that order performances in a small number of categories, such as the four-to six-point range typical of most scoring rubrics. Many scoring rubrics employ at least partly evaluative labels, such as "undeveloped response" or "extensively elaborated response," for the performance categories, as opposed to descriptive labels. In these cases, the rubric embodies a kind of primitive performance standard, at least for the task being evaluated.

The scoring rubric in effect provides a score scale for evaluating task performance and, hence, a basis for setting performance standards for the particular task. At issue is whether the scoring categories have the same meaning across tasks, especially in the face of variations in task difficulty. Whether the same scoring rubric can be meaningfully applied to different tasks to generate a cross-task measurement scale depends on evidence of generalizability of the scoring rubric across tasks. Moreover, because the particular scoring rubric is usually only one among several that might just as well have been formulated, generalizability across scoring rubrics should also be investigated.

These issues of generalizability are especially important for score interpretation and reporting because scoring rubrics are typically task based rather than construct based. That is, more often than not, scoring rubrics refer to aspects of a student's response or product, such as degree of elaboration or coherence, rather than to aspects of process or skill. Going from a task-specific interpretation to a construct interpretation of some generality and power requires evidence of generalizability.

For performance standards to be valid, the increasing achievement levels characterized by such terms as "basic," "proficient," and "advanced"--as well as the tasks that benchmark these levels--should reflect increases in complexity of the construct specified in the content standards and not increasing sources of construct-irrelevant difficulty. However, what constitutes construct-irrelevant variance is a tricky and contentious issue (Messick, 1994a, 1994b). This is especially true of performance assessments, which typically invoke constructs that are higher order and complex in the sense of subsuming or organizing multiple processes.

For example, skill in communicating mathematical ideas might well be considered irrelevant variance in the assessment of mathematical knowledge (although not necessarily vice versa). But both communication skill and mathematical knowledge are considered relevant parts of the higher order construct of mathematical power according to the content standards delineated by the National Council of Teachers of Mathematics. The problem, as was previously mentioned, is to separate evidence of the operation of the focal construct from that of ancillary skills involved in task performance that serve as potential sources of construct-irrelevant difficulty.

The concept of construct-irrelevant variance is important in all educational and psychological measurement, especially in richly contextualized assessments and so-called authentic simulations of real-world tasks, because "paradoxically, the complexity of context is made manageable by contextual clues" (Wiggins, 1993, p. 208). And it matters whether the contextual clues that are responded to are construct relevant or represent construct-irrelevant difficulty. Everything depends on how compelling the evidence and arguments are that the particular source of variance is a relevant part of the focal construct as opposed to affording a plausible rival hypothesis to account for the observed performance regularities and relationships with other variables.

To disentangle construct-relevant from construct-irrelevant variance, one must turn to the construct theory of the subject-matter domain, that is, to the best available integration of scientific evidence about the nature of the domain processes and the ways in which they combine to produce effects or outcomes. A major goal of domain theory is to understand the construct-relevant sources of task difficulty, which then serves as a guide to the rational development and scoring of performance tasks and other assessment formats.

If the theoretical sources of task difficulty are actually used as a guide for test construction, the resulting exercises or tasks should have some critical properties. In particular, their ordering and approximate placement on the measurement scales should be predictable. Empirical evidence that the actual scale placement of these tasks is predicted by theory-based indices of task difficulty then provides strong support for the construct validity of both the theory and the measurement scale--for example, as was done for the prose, document, and quantitative scales in the National Adult Literacy Survey (Kirsch, Jungeblut, & Mosenthal, 1994).

Performance standards are central to standards-based education reform because they are thought to transform educational assessments into worthwhile educational experiences serving to motivate and direct learning. Because performance standards specify the nature and level of knowledge and skill students should attain, the criteria of good performance should become clear to them, in terms of both how the performance is to be scored and what steps might be taken to improve performance. In this sense, the criteria of successful performance are transparent or demystified and hence should be more readily internalized by students as self-directive goals (Baron, 1991; Wiggins, 1993).

Of course, evidence needs to be accrued that the performance standards are understood by students and teachers and that they indeed facilitate learning, because the meaningfulness or transparency of performance standards cannot be taken for granted. In particular, the meaningfulness of the performance standards as applied to the assessment tasks should be appraised. Such evidence bears on the consequential aspect of construct validity.

Also of consequence is the possibility that common performance standards for all students may not uniformly serve as challenges for further growth. For some students they may represent hurdles or artificial barriers that channel educational experiences in ways that are not personally fulfilling, thereby limiting development in line with personal interests and values (Coffman, 1993). Those students who learn different things at different rates may be consigned to failure because they do not learn the common things at the expected rate. Such potential adverse side effects need to be appraised because they bear on the very meaning of the performance standards as well as on their implications for educational policy.

VALIDITY OF STANDARD SETTING

The meaning of content and performance standards also depends in large measure on the credibility of procedures used in setting the standards. Because standard setting inevitably involves human judgment, a central issue is *who* is to make these judgments, that is, whose values are to be embodied in the standards. Thus far, it seems clear that informed judgments about content standards require knowledge of the subject-matter domain as well as of the students' levels of developing expertise. Hence, the group of judges should certainly include teachers and curriculum specialists, who are also appropriate for setting performance standards. An important question in a pluralistic society is who else should participate in the standard-setting process?

The more diverse the group of judges, of course, the less consistency should be expected in their judgments and the more difficulty in reaching consensus. With a heterogeneous group of judges, a range of disagreement around the consensus should be anticipated--disagreement that can be reduced in refined standard-setting procedures by feedback and through discussion among the judges. This range of disagreement has been called a "consensus distribution" (Phillips, Herriot, & Burkett, 1994), which should be robust in being replicable over a variety of settings with the same mix of judges' backgrounds.

An important issue is not just the extensiveness of this distribution, but also whether the distribution represents random variation around the consensus as opposed to consistently different value perspectives or points of view. If the latter is true, some means of accommodating diverse viewpoints needs to be considered to make consensus meaningful under conditions of pluralism (Messick, 1985).

Much of the discussion of the construct validity of performance standards has highlighted the need for convergent and discriminant evidence that supports the meaning of the measurement scale and, in particular, the nature of the cognitive processes entailed at each performance-standard level. Whether the levels themselves are set at the proper points is a contentious issue and depends on the defensibility of the procedures used for determining them. Because setting these levels is inherently judgmental, their validity depends on the reasonableness of the standard-setting process and of its outcome and consequences, not the least of which are passing rates and classification errors.

For example, consider the reasonableness of the widely used Angoff (1971) method of standard setting. In this procedure, expert judges are asked to estimate the probability that a minimally competent respondent will answer each item correctly. The average estimate for each item provides a kind of minimum passing level for the item. These estimates are summed to determine a passing or cut score for the test. Modified versions of the Angoff method are typically used to set nonminimum standards, such as the basic, proficient, and advanced levels of NAEP. The reasonableness of such judgments clearly depends on the expertise of the judges. Therefore, the judges should be knowledgeable not only about the subject-matter domain but also about the performance of persons exhibiting various levels of proficiency in the field.

Other aspects of the reasonableness of this standard-setting process can also be addressed. For example, the logical or internal consistency of the process can be appraised by comparing the Angoff probability estimates for each item with the proportion of minimally competent respondents who get the item correct, such respondents being defined as those scoring at or just above the cut score for the test (Kane, 1984). In one appraisal, the results were modest but encouraging--a correlation of .71

between the mean Angoff probability estimates for the judges and the mean performance of minimally competent respondents (DeMauro & Powers, 1993). However, for correlations between estimated and observed item difficulties, both by item and by judge, medians were in the low 40s, which is consistent with other studies of subject-matter experts' only modest ability to estimate item difficulty and discrimination (Bejar, 1983).

One line of development at this point pursues methods to improve the precision and consistency of the judgmental estimates (e.g., Kane, 1987). Another line might be to identify vulnerabilities in the judgmental process and attempt to overcome them. For example, a major weakness of item-level judgmental procedures such as the Angoff method occurs precisely because judgments are made at the item level for each item separately. When each item is considered in isolation, item-specific variance looms large compared with construct variance. This tends to distort probability estimates that are supposed to reflect levels of *construct* competence. The distortion might be reduced by requiring judgments of the probability of success on small sets of items where the construct variance would be more salient because it cumulates across items while the item-specific variance does not.

Another problem with item-by-item judgments is that they do not capitalize on the structure of interrelations among the items as does IRT scaling or other model-based approaches to developing measurement scales. Indeed, once a measurement scale is constructed, especially if exercises are benchmarked along the scale and validated process descriptions are formulated for various scale levels, standards can be set directly as points on the scale. This involves judgments about what level of process complexity (and of associated scaled exercises) is appropriate to performance at minimal, basic, proficient, or advanced levels.

Another weakness of standards based on item-level methods such as Angoff's is that they may not hold if the items are changed, although extrapolations are generally defensible to equated tests or item sets. In contrast, if standards are set as points on a measurement scale such as those based on Item Response Theory (either directly as just described or by combining judges' probability estimates for items calibrated to the scale), the standards should remain relatively invariant when calibrated items are added or dropped from the set. As a consequence, such scale-level standard setting is amenable to use with computer-adaptive as well as linear tests.

Moreover, if a well-developed theory of the sources of construct-relevant difficulty has guided test construction and if the resultant exercises fall on the scale in their predicted order and approximate expected placement, it may be possible, as was done in the National Adult Literacy Survey (Kirsch et al., 1994), to empirically delineate regions of the scale where construct processes emerge, differentiate, compound, hierarchically integrate, or otherwise become more complex. The empirical delineation of such scale regions then provides a rational basis by which judges can set standards in terms of desired levels of process complexity for different grade levels and degrees of expertise.

Finally, the measurement scale can be elaborated by projecting onto it a variety of other behaviors, scores, and real-world tasks (Messick, Beaton, & Lord, 1983; Phillips, Herriot, & Burkett, 1994). These might include American College Test (ACT), Scholastic Assessment Test (SAT), New York State Regents Examination, or Advanced Placement scores; achievement in math and science; and skill in reading *TV Guide* or *New York Times* editorials as well as high school or college textbooks. In this procedure, which I refer to as behavioral anchoring, *nonassessment* tasks and scores are projected onto the scale. This is in contrast to benchmarking, in which *assessment* tasks mark particular points on the scale.

By adding behavioral anchoring to validated process descriptions, scale levels can be related to a variety of accepted norms, thereby giving policymakers and laypeople alike a better sense of what is implied by the scale levels and hence by standards set as points on the scale. With such information in hand, it becomes possible to open up the standard-setting process beyond the group of experts needed to make item-level judgments (Messick, 1985). By focusing not on isolated items but on the ordered set of benchmark exercises, on the associated process descriptions, and on the implications of behavioral anchoring, meaningful standards judgments could be obtained from policymakers, parents, businesspeople, representatives of minority groups, and other stakeholders in education in a pluralistic society.

OVERVIEW

Because content standards specify what is important to teach and to learn, they provide blueprints for standards-based educational assessment. Because performance standards specify accredited levels of student accomplishment, they require a measurement scale of some type to characterize the location and meaning of those levels. Hence, the validity of content and performance standards cannot be separated from the validity of the assessment itself or of the measurement scale. Therefore, the validity of standards must be addressed in terms of the same criteria needed to appraise the validity of assessments generally. These include content, substantive, structural, generalizability, external, and consequential aspects of construct validity.

With these fundamental aspects of construct validity in mind, the validity of content standards was addressed in terms of their representative coverage of the construct domain and their alignment with students' cognitive levels of developing expertise in the subject matter. The construct validity of performance standards as interpreted score levels was addressed in terms of the extent to which they reflect increasing construct complexity as opposed to construct-irrelevant difficulty.

Operationally, performance standards are interpreted points on a measurement scale. At issue are both the proper placement of those points and their meaning in terms of the knowledge and skill entailed in performance at those levels. The meaning of the performance standards depends on the construct validity of the measurement scale. The appropriateness of their placement depends on the reasonableness of procedures used for setting them. The advantages of standard-setting based on scale-level judgments, as opposed to the compounding of item-level judgments, were explored, especially as they bear on opening up the standard-setting process to a pluralism of stakeholders beyond subject-matter experts.

In sum, it may seem that providing valid grounds for valid inferences in standards-based educational assessment is a costly and complicated enterprise. But when the consequences of the assessment affect accountability decisions and educational policy, this needs to be weighed against the costs of uninformed or invalid inferences.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education, 4,* 305-318.

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7,* 303-310.

Brennan, R. L. (1995). Standard setting from the perspective of generalizability theory. In *Proceedings of Joint Conference on Standard Setting for Large-Scale Assessments.* Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Coffman, W. E. (1993). A king over Egypt, which knew not Joseph. *Educational Measurement: Issues and Practice, 12*(2), 5-8.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity*. Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147-171). Urbana, IL: University of Illinois Press.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

DeMauro, G. E., & Powers, D. E. (1993). Logical consistency of the Angoff method of standard setting (ETS RR 93-26). Princeton, NJ: Educational Testing Service.

Embretson W. S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179-197.

Ferguson, G. A. (1956). On transfer and the abilities of man. *Canadian Journal of Psychology, 10,* 121-131.

Hunter, J. E., Schmidt, F. L., & Jackson, C. B. (1982). *Advanced meta-analysis: Quantitative methods of cumulating research findings across studies*. Newbury, Park, CA: Sage.

Kane, M. (1984, April). *Strategies in validating licensure examinations*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Kane, M. (1987). On the use of IRT models with judgmental standard-setting procedures. *Journal of Educational Measurement, 24,* 333-345.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527-535.

Kirsch, I. S., Jungeblut, A., & Mosenthal, P. B. (1994). Moving toward the measurement of adult literacy. In *Technical report on the 1992 National Adult Literacy Survey*, Washington, DC: U.S. Government Printing Office. Press.

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement, 16,* 294-304.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635-694 [Monograph Supplement 9].

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012-1027.

Messick, S. (1985). Progress toward standards as standards for progress: A potential role for the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice, 4*(4), 16-19.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Messick, S. (1994a). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1994b). *Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning* (ETS RR-94-45). Princeton, NJ: Educational Testing Service.

Messick, S., Beaton, A., & Lord, F. (1983). *National Assessment of Educational Progress reconsidered: A new design for a new era* (NAEP Rep. 83-1). Princeton, NJ: National Assessment of Educational Progress.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., & Phillips, G. W. (1993). *NAEP 1992 mathematics report card for the nation and the states* (Rep. No. 23-ST02). Washington, DC: National Center for Education Statistics.

Phillips, G., Herriot, R., & Burkett, J. (1994). Issues in establishing technical guidelines for standards-based reporting [Draft]. Washington, DC: National Center for Education Statistics.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.) *Review of Research in Education, 19,* (pp. 405-450). Washington, DC: American Educational Research Association.

Shulmann, L. S. (1970). Reconstruction of educational research. *Review of Educational Research, 40,* 371-396.

van der Linden, W. J. (1995). A conceptual analysis of standard setting in large-scale assessments. In *Proceedings of Joint Conference on Standard Setting for Large-Scale Assessments*. Washington, DC: National Assessment Government Board and National Center for Education Statistics.

Wiggins, G. P. (1993, November). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 75*(2), 200-214.

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 75-107). Hillsdale, NJ: Erlbaum.

Standards-Based Score Interpretation: Establishing Valid Grounds for Valid Inferences

Summary of Break-out Session[1]

The discussion focused first on *content standards*. Two initial questions were: (a) How can one evaluate and minimize *slippage* in the judgmental-test-development process? and (b) In standards-based educational assessment, the construct validity of content standards depends on representative *coverage of the subject-matter domain*, but is this in terms of the generic construct specifications, the exercises embodying them, or both? Messick explained that minimizing slippage away from what one intended to measure is a generalizability issue.

Addressing the question related to content coverage, he indicated that test specifications do not cover everything. He explained that committees work to establish the test specifications, which are revised over time. He pointed out that another question would be: How serious is any loss of intended coverage by the set of adequate items? One participant indicated that one way to proceed would be to start with the content standards and performance standards and differentiate basic from proficient items by trying out the items on previously-identified groups of students. Messick indicated that that confounds the issue with irrelevant features; additionally, bias emerges as one goes down to specifics. However, another observed that any real set of tasks are a nonrandom subset of possible tasks that would fit the content and test specifications.

One participant observed that the National Assessment of Educational Progress (NAEP) has only procedural evidence of validity, but that interpretation generalizes to other situations outside of school. The question was: What data do we need to measure validity? Messick referenced the process of developing constructs of sources of difficulty as developed by Mosenthal-Kirsch as one way to collect validity data. One participant pointed out that the Mosenthal-Kirsch theory for item difficulty still has some anomalies in its descriptions; the NAEP reading items do not fit the scale that well. Messick responded that the Mosenthal-Kirsch theory leads to improvements in scale development; he suggested that if the items do not fit the scale, then it is necessary to expand the theory that predicts item difficulty, or else develop the scale items differently. He indicated that the theory says that the more distinctions that need to be made, the more difficult the item, and that the more difficult multiple-choice items include more plausible distractors. Within the group, it was suggested that items are made more difficult because of the use of arcane, esoteric words. At the end of the discussion, Messick indicated that the way to cover the content better is to use more items on the test.

The group gave its attention to another question: How replicable is NAEP as a whole? NAEP is repeated every year or two, but if it were to be replicated immediately, would it find the same distribution of student performance? Messick indicated that it is a good idea to put real-world tasks on the NAEP scales. However, the other participants agreed that it is too costly to completely develop a test repeatedly to check the replicability of the process that produces the test. Additionally, it was noted that even though the test specification committee changes its membership over time, the test continues to fit the specifications over time. It was noted that because the National Assessment Governing Board (NAGB) changes its test specifications over time, this kind of checking is problematic.

---

[1] This is a summary of the break-out discussions for Samuel Messick's presentation. The session was facilitated by Susan Ahmed (National Center for Education Statistics) and recorded by Andrew Kolstad (National Center for Education Statistics).

Other questions were presented to the group as follows: As we move from test specifications to item development, do we need a theoretical basis for developing scoring rubrics across all test items? Is there any theoretical basis for wanting to make the scoring rubric consistent across test items? Messick noted that difficulty gets built into the rubric and changes the meaning of the rubric across different tasks. While some participants indicated that that makes no difference to this kind of standard, one noted that part of the problem depends on the consequences of the classification made.

The discussion then focused on issues of validity. The topic addressed was: NAGB policy defines "basic," "proficient," and "advanced;" the link between educational policy and NAGB cut scores can be documented. Are these two facts evidence for the validity of the cut scores? Messick indicated that there is evidence of the validity of the cut scores provided that the link is more than opinion. He explained that tasks that reflect the policy descriptions can be developed and then evaluated to see how well these tasks fit the NAEP scales. He argued that validity is strengthened by the convergence of several forms of evidence, including external checks. He added that five kinds of validity are needed besides generalizability (the other types are discussed in Messick's paper).

The group then addressed the setting of cut scores and the training of judges. Messick pointed out that the methods used to set cut scores assume that people have some expert knowledge. He recommended that relevant information be provided to those who are not experts, so that they can make better item-level judgments. Within the group, there were varying opinions about the use and training of judges. For example, one participant suggested that judges who are trained assume the expert view of things and lose their "lay" sensibility; another suggested that the use of experts from the general public could make a contribution. Messick emphasized that a broad range of judges is needed to provide legitimacy to the process of setting standards. He added that standards can become invalid if the graduates of the school system perform below what the standards claim they should be able to do.

A series of questions related to the NAEP report were then addressed. These were: (a) Are NAEP reports used? (b) How good are the decisions based on the NAEP reports? (c) Do the NAEP Trial State Assessment (TSA) reports have any impact on state policies? The participants shared their knowledge and individual experiences related to the use of the reports.

The discussion then focused on issues related to adjusting state proficiency scores to account for varying minority composition and socioeconomic status. For example, should the District of Columbia scores be adjusted to reflect the overall United States' racial balance? The group acknowledged that it is necessary to consider the unintended consequences of these adjustments. However, the group suggested some adjustments that the NAEP design could support, as follows: (a) improving *background data* elements, (b) *sampling intact classrooms*, and (c) *the use of classroom variables* for policy analyses. The group also discussed that the much-reported relationship between TV watching and NAEP reading scores is sometimes considered to be a causal relationship, but there are plausible alternatives that could bring about such a relationship without any causality involved; NAEP reporting should offer alternative hypotheses. The group agreed that *a report should be issued that standardizes state scores as a research report* rather than as a mainline NAEP publication.

The group summarized the main points of its discussion as follows: (a) Standard-setting issues are embedded in broader measurement issues, including domain frameworks, test specifications, task development, scoring rubrics, and scaling; and (b) each step in the process requires attention and evaluation. Validity concerns need to be addressed when generalizing across tasks to the domain, when representing the domain with test specifications, and when developing scoring rubrics.

Ensuring Fairness in the Setting of Performance Standards

Lloyd Bond

Professor, School of Education, University of North Carolina at Greensboro

ABSTRACT

Standards of performance on educational measures should distinguish the same levels of knowledge, proficiency, and ability for all test takers. That is, a performance standard that purports to identify "advanced" proficiency on some measure should result in classifications that are equally accurate across all groups of test takers. In like manner, standards that purport to identify those in need of remediation should not result in misclassification rates that vary as a function of ethnicity, gender, or other characteristics that are irrelevant to the construct being assessed.

This presentation discusses (a) methods for investigating the extent to which performance standards may result in differential rates of misclassification of subgroup members in the population of examinees, and (b) procedures for setting initial performance standards that minimize the likelihood of such misclassifications.

INTRODUCTION

The assessment-driven reform under way in education in the United States has caught on faster and garnered more distinguished adherents than perhaps any educational reform in the past 100 years. As recently as 10 years ago, it would have been unthinkable that eminent measurement specialists and policy analysts, chief state school officers, and even local educational agencies would be seriously debating the relative advantages of national content standards and national performance standards over the entrenched local standards that were the cornerstone of American education. The 1983 publication of *A Nation at Risk* is generally taken as the unofficial beginning of the current move toward national performance standards. In that landmark publication, the authors proffered a gloomy diagnosis of the state of American public education and warned of impending doom if radical surgery were not performed. It took a couple of years for the message to sink in, but once it did, the swiftness with which the entrenched, uniquely American aversion to a national curriculum and national standards melted away has been really quite remarkable.

To be sure, the notion that public education can be reformed by assessment-driven initiatives and associated standards of performance is not new. This century has witnessed wave upon wave of educational reform initiatives--many of them test-driven and many having the character of panaceas. In the first quarter of this century, American measurement specialists, foremost among whom was of course Terman, saw the new Binet-Simon test and the paper-and-pencil companion pioneered by Arthur Otis as a boon to American public education. The new tests were seen as a surefire, efficient way to identify the ablest and to weed out the "uneducable." There followed a proliferation of testing in industry, the military, college admissions, and public schools that has continued unabated for almost 100 years. The most recent test-driven initiative, "minimum competency testing," enjoyed a brief and furious popularity but has more or less faded, with contradictory claims as to its overall effect upon student achievement (Darling-Hammond, 1995; Lerner, 1981; Madaus, 1994).

The current initiative, variously called authentic assessment, direct assessment, and performance assessment, may well have a much longer half-life for at least two related reasons. First, unlike the minimum-competency movement, performance assessment attempts to *integrate* teaching and assessment in such a way that the two are complimentary and mutually reinforcing, rather than disjointed. A second reason that performance assessment may be more than just a passing fad is that teachers tend to view it as less threatening and less punitive than the typical, externally imposed standardized test traditionally used for accountability purposes. They have argued, with some justification, that norm-referenced standardized tests used for high-stakes purposes such as student promotion and retention inevitably force them to alter instruction in ways that are educationally unsound.

SETTING PERFORMANCE STANDARDS

A performance standard may be defined as a point on a scale of proficiency that classifies individuals into mutually exclusive categories characterized by certain levels of competence, knowledge, or skill. The categories may or may not carry verbal labels and descriptions, and the interpretations made and actions taken on the basis of the classifications may vary from the relatively innocuous (as when used as convenient summaries to monitor student achievement) to the deadly serious (as when used to place children in classes for the educable, mentally retarded or to deny a high school diploma). The push for national performance standards of proficiency in various academic subjects has generated anew an interest in the technical difficulties in setting such standards and the sociopolitical consequences of using them (Bond, 1993; Darling-Hammond, 1993).

The technical difficulties are many and by no means settled. The essential problem stems from the simple fact that there is no way to set performance standards that enjoys consensus among measurement specialists. A summary of procedures used may be found in Jaeger (1989). The most popular standard-setting methods involve some variation of the Angoff procedure, but these procedures are all currently under a cloud. The National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment concluded that the basic procedure is fatally flawed because it imposes cognitive demands on standard-setting panelists that are impossibly difficult and unrealistic.

The uses to which performance standards are put can have enormous social and political consequences. Standards of performance on various cognitive ability tests often determine the very nature, quality, and course of an individual's early education. In some states, subject-matter achievement tests determine grade-to-grade promotion, receipt of a high school diploma, or admittance to various advanced courses of study. Standards of performance determine who will and will not be allowed to practice many professions and who will and will not be certified in various occupational specialties. Average scores of students in a district or school often are part of a formula that determines allocation of funds for special education.

From a purely technical point of view, the *consequences* of setting a particular performance standard and using it in a particular application can be distinguished from the inherent validity of the classification(s) made on the basis of the standard. In the third edition of *Educational Measurement*, Messick (1989) argued that, broadly conceived, the validity of a test also encompasses the educational, social, and political consequences of its use. This notion has spawned considerable discussion among measurement specialists. Many would prefer to keep the consequences of using a test, important as they are, separate from a formal evaluation of the validity of test interpretations.

ADVERSE IMPACT, BIAS, AND UNFAIRNESS

Any discussion of the equity and fairness implications of performance standards must necessarily address the "fairness" of the assessment itself. That is, it is not possible to set interpretable standards of proficiency on an assessment that is itself "unfair" or biased against specific groups. It may be useful here to distinguish between three related terms: *adverse impact, bias, and unfairness*. An assessment and its associated performance standard are said to exhibit *adverse impact* with respect to race, for example, if the rates at which African American examinees are denied some favored treatment (such as grade promotion, graduation, or employment) are substantially below the rates of majority examinees. The important technical point to note is that *the differential classification rate alone* is sufficient for adverse impact with respect to race to exist. The reason (or reasons) for the differential classification rates is (are) not relevant to a determination of adverse impact. Differential classification rates *may* result from biases in the conceptualization of the domain to be assessed, from characteristics of the scoring scheme that disadvantage African American students, or from unequal opportunity to learn the material and acquire the skills that are assessed. Differential certification rates may not be traceable to any known deficiencies in the assessment system itself, but may represent genuine group differences in the knowledge, skills, and abilities being assessed. Mere presence of adverse impact gives no clue as to which reason or set of reasons is operative. These must be investigated as a separate matter.

An assessment exhibits *bias* against a specific group if significant, systematic differences in the performance of members of the group can be ascribed to actual flaws or deficiencies in one or more aspects of the assessment itself that have the effect of disadvantaging members of that group. Often, such deficiencies in the assessment are due to "construct-irrelevant" factors. For example, in the assessment of mathematical proficiency, it is important to keep to a minimum the level of competence in the language in which the test is written. If the examination includes word problems, then the vocabulary and linguistic demands of the problems should be as simple as possible. Otherwise, persons less proficient in the language of the test, such as those for whom the language is a second language and those who speak specific dialects within the general population, may be penalized because of purely linguistic, as distinct from mathematical, considerations.

It is also important to note that biases may enter an assessment even when the assessment is free of construct-irrelevant factors. For example, assessment exercises may sample the knowledges, abilities, and skills in the content domain in a way that does not reflect the initial content framework. In fact, the content domain itself may be inappropriately specified. It is doubtful whether any current conceptualization of teaching incorporates in the content framework the kinds of skills necessary for even a modicum of effectiveness of some of this nation's big city schools. To take another example, assessments developed for certification in a wide variety of fields often pose vignettes that candidates must critique and evaluate. Typically, the responses entail considerably more writing than would normally be the case in actual performance on the job.

If adverse impact results from biases in the test, then to that extent the test is *unfair*. That is, if use of the test and its associated performance standard results in the denial of some favored treatment to groups because of biases in the test itself, then it is an unfair test.

BEST COPY AVAILABLE

FAIRNESS AND PERFORMANCE ASSESSMENTS

The motivation for a discussion of fairness in setting performance standards on large-scale assessments stems from the historical and pervasive existence of moderate to substantial differences in mean scholastic achievement (as measured by performance on standardized assessments, high school and college graduation rates, and any number of other indicators) among subgroups in the population. African American students and Hispanic students in particular score below white and Asian students on most indicators of educational achievement. The reasons for the lower achievement have been the source of much controversy, with some writers positing a genetic basis for the differences (Jensen, 1980; Murray & Herrnstein, 1994).

One of the many criticisms of multiple-choice tests is that they advantage "test-wise" examinees--those who know how to allocate time optimally, those who know when and when not to guess, those who can detect clues in the stem or distractors of an item that aid in choosing the right answer, and so on. To the extent that test-wiseness contributes nontrivially to performance and that subgroups of the population differ in their level of test sophistication, a reasonable conjecture is that the move away from multiple-choice tests to performance assessment should result in smaller group differences.

In addition to advantaging test-wise students, multiple-choice tests have been criticized on the basis that they trivialize and decontextualize knowledge, that they place undue emphasis on speed, that they penalize deeply reflective students, that they cannot probe deeply into students' understanding, and that they distort instruction. All of these criticisms have an element of truth, but some have been overblown. For example, the charge that multiple-choice tests cannot probe deeply into a student's understanding is probably overstated. The most significant shortcoming of multiple-choice tests, in my opinion, is that the response mode used (blackening ovals) is hopelessly impoverished. No matter how cleverly item writers choose distractors, no strong inference about the student's thinking and the student's state of knowledge can be deduced from a wrong answer. All that is known for certain is that an incorrect answer was chosen. The great promise of performance assessment is that it provides students with alternative ways to demonstrate what they know and that it encourages the occasional creative approach to a problem (perhaps overlooked even by the exercise developer), which becomes a part of the permanent record rather than being ignored by a scanning machine. In the hands of a good teacher, performance assessment encourages precisely the kind of teacher-learner interaction and instructional dialogue that is the basis of sound instruction.

The empirical evidence in support of the notion that performance assessments will result in smaller differences between African American and white examinees is mixed (Baker & O'Neil, in press; Bond, 1993; Darling-Hammond, 1995; Linn, 1993; Linn, Baker, & Dunbar, 1991; Madaus, 1994; National Commission on Testing and Public Policy, 1990; Office of Technology Assessment, 1991; Winfield & Woodard, 1994). Typically, differences in performance between groups are compared for traditional and alternative assessments in terms of effect size or by comparing the proportions scoring above or below some cut score. Effect sizes are computed by subtracting the mean for the minority group of concern from the mean for the majority comparison group and dividing by the standard deviation of the majority comparison group. A narrowing of the "achievement gap" would be indicated by a smaller effect size or greater equivalence across groups in proportions passing when performance assessments are compared to more traditional assessments.

Some studies have found the achievement gap between African American and white students on performance assessments is comparable to the gap found on multiple-choice measures of the same or similar constructs. For example, Linn, Baker, and Dunbar (1991) found that differences in achievement between African American and white students in grades 4, 8, and 12 on the 1988 NAEP were essentially the same for the open-ended essay items in the writing assessment and for the primarily multiple-choice items in the reading assessment. Elliott (1993), when comparing differences in performance for white and Asian students with performance differences for African American and Latino students on multiple-choice and constructed-response items for the 1992 NAEP in mathematics, found a larger gap for constructed-response items. LeMahieu (1992, cited in Madaus, (1994)) reported larger differences between African American and white students when portfolio scores, reflecting self-selected pieces of writing, were compared to independent measures of writing. LeMahieu attributed the differences to self-selection: African American students did not choose material from their writing folders that best represented their writing.

By contrast, Badger (1995), using results of open-ended and multiple-choice items from the Massachusetts testing program, found smaller performance gaps for open-ended items when comparing students in schools with low and high Socioeconomic Status (SES), and when comparing African American and Latino students to white and Asian students. These results were consistent across grades and subject areas. Johnson (1988) compared passing rates for African American and white teachers on the multiple-choice National Teacher Examination (NTE) with passing rates for classroom observations and found that the differences observed across groups on the NTE disappeared in the classroom observation. A comparison of the magnitude of differences between African American and white military personnel on paper-and-pencil tests (Armed Services Vocational Aptitude Battery [ASVAB]) and job knowledge tests versus hands-on tests resulted in substantially smaller differences for the hands-on tests (Wigdor & Green, 1991).

It is difficult to draw definitive conclusions from the above studies. The studies differed widely in the constructs assessed, the purposes of the assessment, the format in which performances were presented and evaluated, the antecedent instructional conditions, and the social and academic characteristics of the students assessed. There is no particular reason to suppose that all of the educational disadvantages known to characterize many urban schools and many urban communities will disappear if the forms of assessment are changed. Nevertheless, the results of Johnson (1988) and Wigdor and Green (1991) suggest subgroups' differences in real-world, hands-on performance are smaller than those found in more academic tasks.

Bond, Moss, and Carr (in press) emphasize that careful attention must be given to the way in which group differences are reported and interpreted. Reporting simple differences in performance between schools or districts, between racial/ethnic groups, between students from poor families and those from wealthy families, without additional information to assist in *explaining* the differences, may result in serious misinterpretations. For instance, changes in assessment scores from year to year may simply reflect changes in the student population (dropouts or transfers, for example) rather than changes in the capabilities of students. Differences in assessment scores across ethnic groups in part reflect differences in socioeconomic status of the communities in which the groups live. Differences in assessment scores from school to school in part reflect differences in resources and programs such as the qualification of teachers or the number of advanced course offerings. Racial breakdowns by themselves overestimate the importance of racial differences in academic achievement because such differences are confounded by uncontrolled socioeconomic differences. This risks misinforming the nation by allowing people to conclude that it is only race (and not poverty) that is driving these differences. Differences in achievement between underserved minority and majority groups (whether defined by race/ethnicity or socioeconomic status) are reduced when one controls for socioeconomic status and can be reduced even further when one controls for differences in access to high-quality education (Dreeben & Barr, 1983; Darling-Hammond, 1995;). Comparisons of test scores that ignore these factors hold little promise of informing policy or of directing policymakers' attention to the real sources of the problem.

FAIRNESS AND THE SETTING OF PERFORMANCE STANDARDS

The setting of a performance standard on a given test implies that a more or less unitary construct is being measured. This is one reason why the notion that a given total score on the NAEP eighth-grade mathematics test, say, represents some specific level of math knowledge and that understanding is problematic. The NAEP eighth-grade math assessment simultaneously reflects a student's knowledge of numbers and operations, algebra, geometry, and measurement and statistics as well as that student's conceptual understanding, procedural knowledge, and computational skill. The test is a heterogeneous mixture of knowledge and skill, and it is entirely possible that two individuals with quite different patterns of ability will obtain the same total score.

Setting standards of proficiency on performance-based assessments presents technical difficulties of interpretation that are even more problematic. Often, performance-based assessments are deliberately designed to measure complex constructs. Students are allowed considerable latitude in how they demonstrate their knowledge and understanding. The scoring of performance assessments is a decidedly nonroutine enterprise requiring intimate knowledge of the subject matter domain and, in some cases, intimate knowledge of the individual student's instructional background.

Assuming that these conceptual and technical difficulties can be overcome, several principles should guide the setting of performance standards so as to ensure fairness to all examinees. The first principle concerns who specifies the content of the assessment and who sets the actual performance standards. A cardinal rule in the specification of the content domain for an assessment and in the setting of performance standards is that those who set the standards should be thoroughly knowledgeable about the content domain that is to be assessed, the population of examinees who are to take the assessment, and the uses to which the results are to be put. The most logical choice for individuals responsible for specifying the content domain for what students at various grade levels should know and be able to do are outstanding teachers of the various grade levels and subject matters assessed. Accomplished teachers should also be intimately involved in the setting of standards of proficiency. It is not entirely clear what other members of society (such as political leaders, representatives from industry, or parents) should be involved in this process because few adults other than teachers have an adequate knowledge of what abilities and knowledges are reasonable at various age levels.

In the context of the present discussion, biases may enter into the performance-setting process from the very beginning, that is, from the initial conceptualization of the content domain to be assessed. Although it may be relatively easy to get consensus on the content of a math assessment and the level of proficiency required at various grade levels, a similarly easy consensus regarding the content of an assessment of history and social studies is far less likely. One fact seems certain: A panel selected to specify the domain of content in areas such as history and social studies that is composed exclusively of, say, white females who teach in suburban schools is unlikely to be credible to either the public or the teaching profession. The same applies with equal force to any hypothetical panel composed exclusively and homogeneously of members of distinct social, ethnic, or gender groups. Diversity of *perspective* should take precedence over ethnic, gender, or "cultural" diversity, per se, although it is unlikely that the former can be completely assured without some attention to the latter.

The second principle has to do with student assessment that is tied to high school graduation or some other high-stakes purpose such as placement in remedial education or eligibility for advanced courses. In using performance standards to award or deny a high school diploma, ensuring that all affected students have had an opportunity to learn the material covered is now, of course, a matter of legal precedent (*Debra P. v. Turlington*, 1979). Bond, Moss, and Carr (in press) note that obtaining information on students' opportunity to learn is also essential for understanding the assessment results and for directing attention to needed reforms in policy and practice. They note that assessments not only document the success of learning opportunities, but they constrain and enable future learning opportunities as well. Thus, information on the quality and quantity of students' prior instruction in a discipline is legally required when tests are used in high-stakes situations, and such information is instructionally useful.

A third principle, closely related to the first two, is that proficiency classifications should have the same meaning for all subgroups. Essentially, this principle says that the same constructs should be assessed in all subgroups of the test-taking population. It has already been mentioned that problems arise when attempting to assess proficiency in mathematics if students differ in their level of proficiency in the language in which the assessment is written. On first glance, this principle appears to be a strictly psychometric issue, but in fact, it has significant policy implications as well. To illustrate this point, consider two hypothetical groups of students, both of whom are held to the same standard of proficiency and knowledge of, say, basic biological and physical science. For the first group, the district or state requires relevant courses in biology and physical science, and the students have had the benefit of excellent instruction and superior laboratory facilities. For the second group, however, there is no explicit requirement for relevant courses in the sciences, and the students have not taken such courses. Predictably, students in the first group outperformed those in the second group. Properly interpreted, the result could be taken as evidence of the construct validity of test. In fact, the above situation is a demonstration of the method of "known" or "contrasted" groups, an old and venerable construct validation procedure in psychology and education.

Now suppose that both groups of students were required to take relevant science courses, but that the quality of instruction and the quality of laboratory facilities for one group were decidedly superior to those of the other group. Although an examination of their transcripts would not reveal it, this is again an illustration of the known-groups methodology, and it could be argued with equal cogency that one group did not have a fair opportunity to learn the material assessed. Note that in neither of these situations is the assessment itself deficient. Rather, the equity concern goes to matters of interpretation and policy, rather than to quality of assessment.

A fourth principle requires that if the assessment is used as a screen for future educational opportunities, the content of the assessment and the level of proficiency required should be demonstrably related to future success. This principle appears so logical and sensible that one may wonder why it is even included here. It is included because the principle is so often violated. Perhaps the most blatant example is the performance standards once used for pilot selection in the military. Even if women had not been overtly barred from becoming pilots, physical strength standards and other requirements of dubious relation to pilot effectiveness would have barred the vast majority of them anyway. Similar practices still apply in the admissions policies of many professional schools. The number of practicing physicians, for example, who find any use for calculus in their work can be counted on one hand. But many medical admissions officials are still mightily impressed with high-level math courses on a grade transcript. The usual argument is that it is not calculus, per se, that is necessary to become a competent physician, but the habits of mind that it implies. This argument simply is not, nor has it ever been, convincing. The fact that many medical schools no longer require such a rigorous physical sciences background, with no deterioration whatever in the quality of the doctors they graduate, is evidence that the performance standard was unrealistic in the first place.

Finally, as Messick (1989) and others have stressed, more attention must be paid to the *consequences* of particular uses of an assessment. Issues of test fairness properly extend beyond an analysis of the internal properties of the assessment itself to the uses to which tests are put and the social and educational consequences of using a particular test and a particular performance standard. Shepard (1989) has argued forcefully that students assigned to lower and slower educational tracks, usually on the basis of a test, are not well served thereby and would be better off educationally in classes that allow them to demonstrate what they in fact can do, rather than having their educational experiences limited on the basis of tests that purport to reveal what they cannot do. Assessment results should not foreclose additional learning opportunities for students or consign them to permanent, dead-end educational tracks.

THE PROMISE OF PERFORMANCE ASSESSMENT

Numerous surveys taken over the past 10 years indicate that most people in this society believe that if people with congenital physical defects are excluded, all students, regardless of race, gender, or social class, should be held to some reasonable standard of proficiency regarding fundamental competencies for citizenship and the adult world of work. In the mathematics education community in particular, the view that all students are capable of learning and using math has changed in fundamental ways. It was once thought that higher order thinking in math was the special preserve of students in advanced classes. The current view is that each student can benefit from and should have access to a math education that includes encouraging higher order thinking and learning of important mathematics.

To the extent that performance assessments result in student productions that reflect more closely what it is society wants students to know and be able to do, they hold the promise of reversing society's fascination with tests, per se, and the seeming precision with which individuals can be rank ordered with a single set of numbers. The fascination with quantification, by the way, is compounded with an equal fascination, if not fear, of computers. A former mentor of mine tells the story of a colleague who taught a required psychology course at a university full of highly competitive pre-med students. After each examination was scored and returned, a long line of disgruntled students would form outside the professor's office to complain about the grade they received. A fellow faculty member suggested that, instead of handwriting the grades, the instructor should type the grades into the computer and have the computer display the grades. Even though the computer had no hand in determining the grades, but simply spewed out what was typed in, the number of complaints dropped dramatically. There is no known antidote to this fascination with numbers and computers.

It is also the case that too often in the past, the use of tests in education and employment have been instances of the "tail wagging the dog." Many graduate schools, for example, have a rigid policy that, to be admitted, all students must take a standardized admissions test. Even when students have taken graduate courses and performed brilliantly in them, if they want to be formally admitted, they must go back and take the GRE, the LSAT, or some other test required by the program. Such practices indicate a profoundly misguided notion of what tests are all about and their proper place in educational decision making. McClelland (1973), in an article entitled "Testing for competence rather than 'intelligence'" (an article that foreshadowed much of the thinking about performance assessment), relates the story of an

Africian American student applicant for graduate school at Harvard who scored extremely low on the Miller Analogies Test (MAT), but who obviously could write clearly and effectively as shown by the stories he had written for his college newspaper. Despite the fact that the student had demonstrated the very ability the MAT is designed to predict, McClelland could not convince his colleagues to accept the student. He wrote, "It is amazing to me how often my colleagues say things like: 'I don't care how well he can write. Just look at those test scores'" (p. 10).

As stated earlier, national standards of educational achievement, particularly when social rewards (such as grade promotion, high school graduation, or admission to college) are attached, are antithetical to considerations of fairness and equity if there is substantial inequality in educational opportunity among the population of affected students. Having said that, it must be admitted that some degree of inequality of educational opportunity is probably inevitable. There are factors affecting equality of educational opportunity--poverty, racism, broken homes, and despair--that cannot be changed substantially, or even appreciably, by educational testing reforms. The goal should be to remove any *official* barriers to educational opportunity for all people and to encourage universal acceptance of the fundamental premise that all children can learn.

References

Badger, E. (1995). The effect of expectations on achieving equity in state-wide testing: Lessons from Massachusetts. In M. T. Nettles and A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 289-308). Boston: Kluwer Academic.

Baker, E. L., & O'Neil, H. F., Jr. (1995). Diversity, assessment, and equity in educational reform. In M. T. Nettles and A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 69-87). Boston: Kluwer Academic.

Bond, L. (1993). Making innovative assessments fair and valid. In *What we can learn from performance assessment for the professions*. Proceedings of the 1993 ETS Invitational Conference. Princeton, NJ: Educational Testing Service.

Bond, L., Moss, P. A., & Carr, M. (in press). Fairness, equity, and bias in performance assessment. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment*. Washington, DC: National Center for Education Statistics.

Darling-Hammond, L. (1995). Equity issues in performance-based assessment. In M. T. Nettles and A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 89-114). Boston: Kluwer Academic.

Debra P. v. Turlington, 474 F. Supp. 244 (M.D. Fla. 1979), *aff'd in part, rev'd in part,* 644 F.2d 397 (5th Cir. 1981); *on remand,* 564 F. Supp. 177 (M.D. Fla. 1983), *aff'd,* 730 F.2d 1405 (11th Cir. 1984).

Dreeban, R., & Barr, R. (1983). *Educational policy and the working of schools.* New York: Longman.

Elliott, E. (1993). *National testing and assessment strategies: Equity implications of leading proposals for national examinations.* Paper presented at the Ford Foundation Symposium on Equity and Educational Testing and Assessment, Washington, DC.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). New York: American Council on Education/MacMillian.

Jensen, R. (1980). *Bias in mental testing*. New York: Free Press.

Johnson, S. (1988). Validity and bias in teacher certification testing. In R. G. Allan, P. M. Nassif, and S. M. Elliot (Eds.), *Bias issues in teacher certification testing*. Hillsdale, NJ: Erlbaum.

Lerner, B. (1981). The minimum competency testing movement: Social, scientific, and legal implications. *American Psychologist, 36*(10), 1057-1066.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*(1), 1-16.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Madaus, G. F. (1993, October). *Assessment issues around the re-authorization of Chapter 1*. Paper presented to the National Academy of Education, University of Michigan, School of Education, Ann Arbor.

Madaus, G. F. (1994). A technological and historical consideration of equity issues associated with proposals to change the nation's testing policy. *Harvard Educational Review,* 64(1), 76-95.

McClelland, D. C. (1973). Testing for competence rather than "intelligence." *The American Psychologist, 28*(1), 1-14.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Murray, C., & Herrnstein, R. J. (1994). *The bell curve*. New York: Free Press.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Government Printing Office.

Shepard, L. A. (1989). Identification of mild handicaps. In R. L. Linn (Ed.), *Educational measurement,* (3rd ed., pp. 545-572). New York: American Council on Education/MacMillan.

Wigdor, A. K., & Green, B. F. (1991). (Eds.). *Performance assessment for the workplace* (Vol.I). Washington, DC: National Academy Press.

Ensuring Fairness in the Setting of Performance Standards

Summary of Break-out Session[1]

The group began its discussion with the topic: *different standards for different groups*. The participants acknowledged that consideration must be given to the characteristics of many different categories of student groups when setting standards for assessments. It was suggested that while on the one hand it is unfair to promote lesser standards for racial and ethnic minority groups or for schools in economically depressed areas, some consideration should be given to the interpretation of the results of the assessment and the decisions made as a consequence of those results.

The discussion focused on *producing unbiased standards*. It was noted that bias is a matter of test design and construction as much as standard setting. It was suggested further that the composition of the judges' panel affects the standard-setting process, as for example, when the composition of the group reflects sensitivity to cultural and language minorities. The group suggested that research is needed on, first, the differential impact of mode of assessment on test bias and standard setting, and, second, on how weighting parts of tests can serve to limit bias.

The discussion then focused on the distinction between fairness and bias (and adverse impact). One participant suggested that there is only a nuanced difference between "fairness" and "bias"--they are both sides of the same coin. It was noted that the purposes and uses of tests are core bases for determining fairness and validity of the tests; these purposes and uses must be made clear to judges in the standard-setting process.

In a discussion of *controlling and designing the development of the standard-setting process to limit bias*, it was suggested that bias should be controlled from the beginning of the development process so that unbiased standards are not put onto a biased instrument. It was suggested that when the total test development process is done with maximum proficiency, then unfairness is less likely and the process limits the adverse impact of possible bias in standards.

The discussion then turned to the *implications for bias in the National Assessment of Educational Progress (NAEP) standard setting related to the absence of individual results or decisions*. Here it was noted that in NAEP, there is sensitivity to minority concerns in item development; however, there is not that same concern in the selection of the panel of judges where there is little minority participation. One group member stated emphatically that there is no bias in NAEP standards, because no individual student results are reported. Other group members contested this position, insisting that while there are no individual results, individuals are affected by NAEP results. Further, it was argued that as the "nation's report card," NAEP has an ethical responsibility to do a fair assessment; this responsibility extends beyond technical correctness, to reporting and data presentation, as well as, to the education of the public and the media. It was suggested that the potential for adverse impact will always be present as long as performance levels are reported in terms of distribution above and below the standards.

---

[1] This is a summary of the break-out discussions for Lloyd Bond's presentation. The session was facilitated by Stephen Lazer (Educational Testing Service) and recorded by Patricia Dabbs (National Center for Education Statistics).

With regard to *equity concerns related to unit of analysis (for example, the student, school, state) and interpretation of results*, it is important not to ignore the consequences of standards. One participant observed that it is a mistake to treat *low-stakes and high-stakes assessments* in the same way. Another suggested that sometimes the distinction between low stakes and high stakes becomes blurred in practice, e.g., any school site at the 25th percentile rank in the state is labelled an at-risk school.

The participants also discussed the *use of opportunity to learn (OTL) data in the standard-setting process*. It was suggested that there are standard-setting situations where OTL information is not relevant. It was recommended that research is needed on integrating OTL standards into the standard-setting process for assessment where necessary. The group questioned if data should be adjusted based on varying OTL in different areas.

Using Performance Standards to Report National and State Assessment Data: Are the Reports Understandable and How Can They Be Improved?

Ronald K. Hambleton, Professor of Education, and Sharon Slater, Graduate Student
University of Massachusetts at Amherst

ABSTRACT

Considerable evidence suggests that policymakers, educators, the media, and the public do not understand national and state test results. The problems appear to be twofold: the scales on which scores are reported are confusing, and the report forms themselves are often too complex for the intended audiences.

This paper addresses two topics. The first is test score reporting scales and how to make them more meaningful for policymakers, educators, and the media. Of special importance is the use of performance standards in score reporting. The second topic is the actual report forms that communicate results to policymakers, educators, and the public. Using some results from a 1994 interview study with 60 participants using the Executive Summary from the 1992 National Assessment of Educational Progress mathematics assessment, the paper highlights problems in score reporting and suggests guidelines for improved score reporting.

Consider these two quotes from newspaper stories about the 1990 National Assessment of Educational Progress (NAEP) national and state mathematics results:

> Just one in seven eighth grade students nationwide can exhibit average proficiency in mathematics.
>
> Standardized tests of student achievement have shown a peculiar quirk for some time now: Every state's kids somehow manage to score above average. Now at least we've got something different--a national math test in which every state's kids scored below average.

The writer of the first story needs a lesson in basic statistics. If some students score below the average, then other students must score above it. This same writer also confuses the NAEP scale for reporting proficiency scores with the category or interval on the NAEP scale associated with being "Proficient" (other intervals exist for "Below Basic," "Basic," and "Advanced"). These categories or intervals are defined by the performance standards on the NAEP scale.

As for the second quote, again, the writer would benefit from a lesson in statistics. Every student on an NAEP assessment, or any other test, cannot be below average no matter how poor the overall group performance.

Clearly, both quotes are misstatements of the actual NAEP results. Perhaps they were made to help sell newspapers. A more likely explanation in these two instances is that the writers did not understand the NAEP reporting scale or the performance standards that were set on the NAEP scale to aid in reporting results. The latter explanation seems plausible because there are still many people who are unable to distinguish percentages from percentiles, and who believe that a score of 70 on an IQ test is passing and 100 is perfect. Examples of misinterpretations of educational and psychological test scores abound.

Beyond whatever basic quantitative literacy may be lacking on the part of educators, policymakers, and the press, with respect to understanding assessment results, the fact is that interpreting test scores will always be a cognitively challenging activity. Not only do the reporting scales themselves vary from one test to the next, but also both measurement and sampling errors must be considered in interpreting scores. Add performance standards to the scale and the task of interpreting scores becomes even more complex.

This paper addresses two topics related to the use of performance standards in score reporting. The first is test score reporting scales and how to make them more meaningful for policymakers, educators, and the media. Of special importance is the use of performance standards or standards (or achievement levels, as they are called by the National Assessment Governing Board [NAGB]). The second topic is the actual report forms that communicate results to policymakers, educators, and the public. A few results from a study in which we investigated the understandability of an NAEP report illustrate several problems encountered by policymakers, educators, and the media.

REPORTING SCALES

What in the world does an NAEP score of 220 mean? This was a common question asked by policymakers, educators, and the media in a study we conducted in 1994 using the Executive Summary of the 1992 NAEP national and state mathematics results. It is also asked by people attempting to make sense of intelligence (IQ), Scholastic Assessment Test (SAT), American College Test (ACT), and NAEP scores. The fact is that people are more familiar with popular ratio scales, such as those used in measuring distance, time, and weight, than they are with educational and psychological test score scales. Even the thermometer scale, which is an equal interval scale, has meaningful numbers on it to help users understand and interpret temperature scores when they need to. These include 32 and 68, as well as daily experiences (such as yesterday's temperature).

In contrast, test scores are much more elusive. Even the popular percent score scale, which many people think they understand, is not useful unless (a) the domain of content to which percent scores are referenced is clear, and (b) the method used for selecting assessment items is known.

One solution to the score interpretation problem is simply to interpret the scores in a normative way--that is, scores obtain meaning or interpretability by being referenced to a well-defined norm group. All of the popular norm-referenced tests use norms to assist in test score interpretations. However, many state and national assessments are criterion-referenced assessments, and with these assessments, scores need to be interpreted in relation to content domains, anchor points, and performance standards (Hambleton, 1994).

With NAEP, an arbitrary scale was constructed with scores in theory ranging from 0 to 500 for each subject area. Basically, the scale was obtained in the following way: The distributions of scores from nationally representative samples of 4th-, 8th-, and 12th-grade students were combined and scaled to a mean of 250 and a standard deviation of about 50 (Beaton & Johnson, 1992). The task then was to facilitate criterion-referenced score interpretations on this scale (see, for example, Phillips et al., 1993). Placing benchmarks such as grade-level means, state means, and performance of various subgroups of students (such as males, females, Blacks, Hispanics) is helpful in bringing meaning to the scale, but these benchmarks provide only a norm-referenced basis for score interpretations.

One way to make statistical results more meaningful to intended audiences is to report the results by connecting them to numbers that may be better understood than test scores and test score scales. For example, in 1994, when the Federal Aviation Administration (FAA) wanted to calm the public's fears about flight safety, it reported that, with current safety records, a person could fly every day for the next 28,000 years without being involved in a serious flight mishap. By connecting accident rates and the number of years an individual could travel without being in an accident, the FAA probably made the numbers more meaningful for many people and helped them better understand the current record of flight safety.

Table 1 displays some NAEP scores for students at the 10th, 50th, and 90th percentiles on the 1992 mathematics assessment in grades 4, 8, and 12. One of the reported results was that the average 8th-grade student in 1992 performed 5 points higher (i.e., better) than the average 8th-grade student in 1990 (Mullis, Dossey, Owen, & Phillips, 1993). It is clear from Table 1 that the typical student (i.e., the student at the 50th percentile) between 4th and 8th grade gained about 48 points, which converts to about 1.2 points per month of instruction (a gain of 48 points over 40 months of instruction). Recognizing that the growth over the four years is not necessarily linear (see, for example, grade-equivalent scores on standardized achievement tests), it might be said that a gain of 5 points, is roughly equivalent to about 6 months of regular classroom instruction (5 points x 1.2 points gain per month) between grade 4 and grade 8. A 5-point gain in mathematics achievement for the average student moving between the 4th and 8th grades is very sizable and practically significant, and this point would be clear to most people if the gains were reported in terms of months of instruction required to achieve the gain. Using Table 1, similar interpretations could be set up for low- and high-achieving students (i.e., students at the 10th and 90th percentiles of the score distribution) between grades 4 and 8, as well as for those between grades 8 and 12.

Reporting score gains in terms of equivalent months of instruction is one convenient way for audiences to have an understanding of the meaning of NAEP scores and gains in achievement. We have been asked (Skip Livingston during the discussion portion of our session at the National Center for Education Statistics [NCES]-NAGB conference) if we have not simply reinvented the unpopular and commonly misinterpreted grade-equivalent scores. Certainly, we are using the grade-equivalent score concept. But, because we are not reporting scores for individual students, and given the way in which we use grade-equivalent scores in our approach, we have not encountered most of the well-known shortcomings of grade-equivalent scores. The main advantage of our approach (i.e., improved communication of the meaning of NAEP scores and gains) seems to far outweigh any disadvantages of this approach to interpreting NAEP scores.

Other possibilities of considerable promise for criterion-referenced interpretations of scores include anchor points and performance standards (see Phillips et al., 1993). Both of these, however, have caused controversy and debate in the measurement community (see Forsyth, 1991; Shepard, Glaser, Linn, & Bohrnstedt, 1993; Stufflebeam, Jaeger, & Scriven, 1991).

BEST COPY AVAILABLE

Table 1

1992 National Assessment of Educational Progress Mathematics Results

Grade	Percentile Points		
	P_{10}	P_{50}	P_{90}
4	175	220	259
8	220	268	315
12	253	300	343

Both anchor points and performance standards capitalize on the fact that Item Response Theory (IRT)-based scales locate both the assessment material and the examinees on the same reporting scale. Thus, at any particular point of interest (i.e., ability level), the sorts of tasks that examinees can handle can be described. And, if of interest, tasks that these examinees cannot handle with some stated degree of accuracy (e.g., 50% probability of successful completion) can be identified. Descriptions at these points of interest can be developed, and exemplary items can also be selected--that is, items can be selected to highlight what examinees at these points of interest might be expected to be able to do (see Mullis, 1991).

Figure 1 shows the "item characteristic curves" for two NAEP items (see Hambleton, Swaminathan, & Rogers, 1991). At any point on the NAEP achievement (i.e., proficiency) scale, the probability of a correct response (i.e., answer) can be determined. Item 2 is the more difficult item because, regardless of ability, the probability of a correct response to Item 2 is lower than for Item 1. The ability at which an examinee has an 80% probability of success on an item is called the "RP80" for the item. In Figure 1, it can be estimated that the RP80 for Item 1 is about 210 and the RP80 for Item 2 is about 306. This is known as "item mapping," in that each item in NAEP is located on the NAEP achievement scale according to RP80 values. If 80% probability is defined as the probability at which an examinee can reasonably be expected to know something or be able to do something (and other probabilities have often been used, say 65%, with the corresponding RP65 values), then an examinee with an ability score of, say, 210, could be expected to answer items such as Item 1 and other items with RP80 values of about 210 on a fairly consistent basis (i.e., about 80% of the time). In this way, a limited type of criterion-referenced interpretation can be made even though examinees with scores of about 210 may never have actually been administered Item 1 or other items like it as part of their assessment.

The validity of the criterion-referenced interpretations depends on the extent to which a unidimensional reporting scale fits the data to which it is applied. If a group of examinees scores at, say, 270, then a score of 270 can be made meaningful by describing the contents of items like those with RP80 values of about 270. The item-mapping method is one way to facilitate criterion-referenced interpretations of points on the NAEP scale or any other scale to which items have been referenced. Cautions with this approach have been clearly outlined by Forsyth (1991). One of the main concerns has to do with the nature of the inferences that can legitimately be made from predicted performance on a few test items.

A variation on the item-mapping method is to select arbitrary points on a scale and then to thoroughly describe these points via the knowledge and skills measured by items with RP80 values in the neighborhood of the selected points. In the case of NAEP reporting, arbitrarily selected points have been 150, 200, 250, 300, and 350. Then the item-mapping method can be used to select items that can be answered correctly by examinees at those points. For example, using the item-characteristic curves reported in Figure 2, at 200, items such as Items 1 and 2 might be selected. At 250, Item 4 would be selected. At 300, Items 4 and 5 would be selected. At 350, Item 6 would be selected. Of course, in practice there may be many items available for making selections to describe the knowledge and skills associated with performance at particular points along the ability scale. With NAEP currently, RP65 values rather than RP80 values are used; in addition, items that clearly distinguish between anchor points are preferred when describing anchor points. (For more details on current practices, see Mullis, 1991; Phillips et al., 1993; and Beaton & Allen, 1992.)

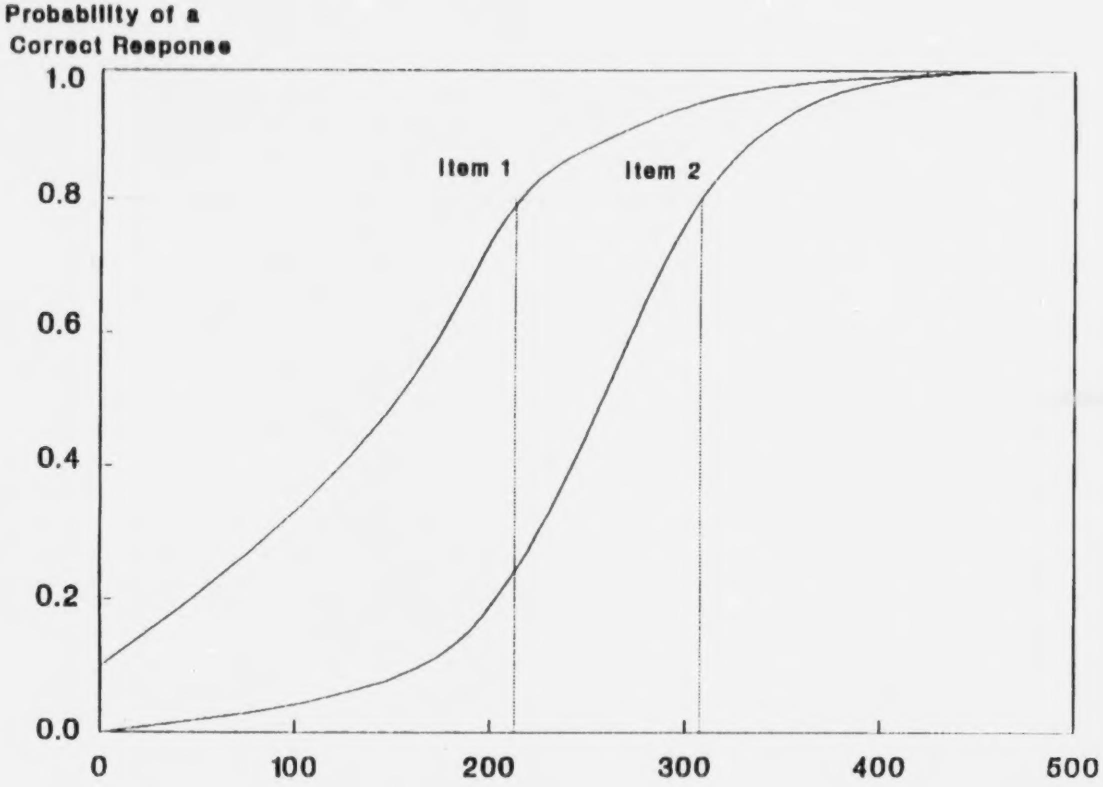


Figure 1. Two test items (or tasks) located on the NAEP achievement scale.

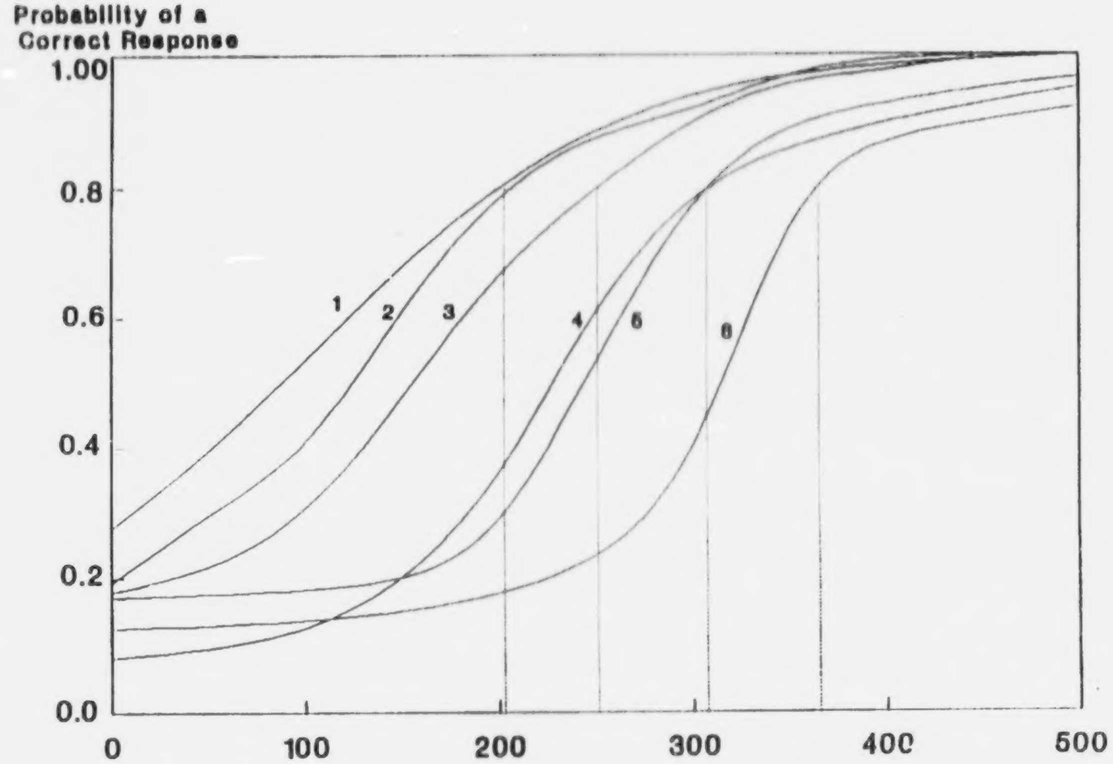


Figure 2. Using anchor points to increase the meaningfulness of NAEP achievement scores.

The NAGB was not completely happy with the use of arbitrary points (i.e., anchor points) for reporting NAEP results. One reason was that the points 200, 250, and 300 became incorrectly associated by the media and policymakers with the standards of performance demanded of 4th-, 8th-, and 12th-grade students, respectively. To eliminate the confusion, as well as to respond to the demand from some policymakers and educators for real performance standards on the NAEP scale, NAGB initiated a project to establish performance standards on the 1990 NAEP mathematics assessment (Bourque & Garrison, 1991; Hambleton & Bourque, 1991) and conducted similar projects to set performance standards on NAEP assessments in 1992 and 1994. The standards have been controversial (see, for example, American College Testing, 1993; Shepard, Glaser, Linn, & Bohrnstedt, 1993; Stufflebeam et al., 1991), but that topic will not be taken up here. The important point is that the performance standards provide, to the extent that validity evidence supports their use, an additional basis for interpreting scores within a criterion-referenced framework.

Figure 3 depicts, basically, the way in which performance standards (set on the test score metric, a scale that is more familiar to panelists setting standards than the NAEP achievement scale) are mapped or placed on to the NAEP achievement scale using the "test characteristic curve" (TCC). (The TCC is a weighted-average, item-characteristic curve for the items that make up the assessment.) With the performance standards for a particular grade on the NAEP achievement scale, these standards can be used to report and interpret the actual performance of the national sample or of any subgroup of interest. This situation is represented in Figure 4. With the performance standards in place, the percentage of students in each of the performance categories in score distributions of interest can be reported, and the changes in these percentages can be monitored over time.

Anchor points and performance standards are placed on an achievement scale to enhance the content meaning of scores and to facilitate meaningful criterion-referenced interpretations of the results (e.g., What percentage of grade 4 students in 1992 was able to perform at the Proficient level or above?). In NAEP reporting in recent years, both anchor points (e.g., 150, 200, 250, 300, and 350) and performance standards (e.g., borderline scores for Basic, Proficient, and Advanced students at grades 4, 8, and 12) have been placed on these NAEP scales. Many states have adopted similar techniques for score reporting.

Performance standards are more problematic than anchor points because they require a fairly elaborate process to establish (e.g., 20 panelists working for 5 days at a grade level) and validate. At the same time, performance standards appear to be greatly valued by many policymakers and educators. For example, many state departments of education use performance standards in reporting, (e.g., Kentucky) and many states involved in the NAEP trial state assessment have indicated a strong preference for standards-based reporting over the use of anchor points.

STANDARDS-BASED REPORTING

Performance standards can provide a useful frame of reference for interpreting test score data such as NAEP. And, with respect to NAEP, policymakers, educators, the media, and the public need a frame of reference to make sense of the plethora of statistical information coming from 25 years of national assessments. Scaled scores without performance standards (or anchors) would convey little meaning to anyone. But it is not enough to have defensible and valid performance standards. They must be

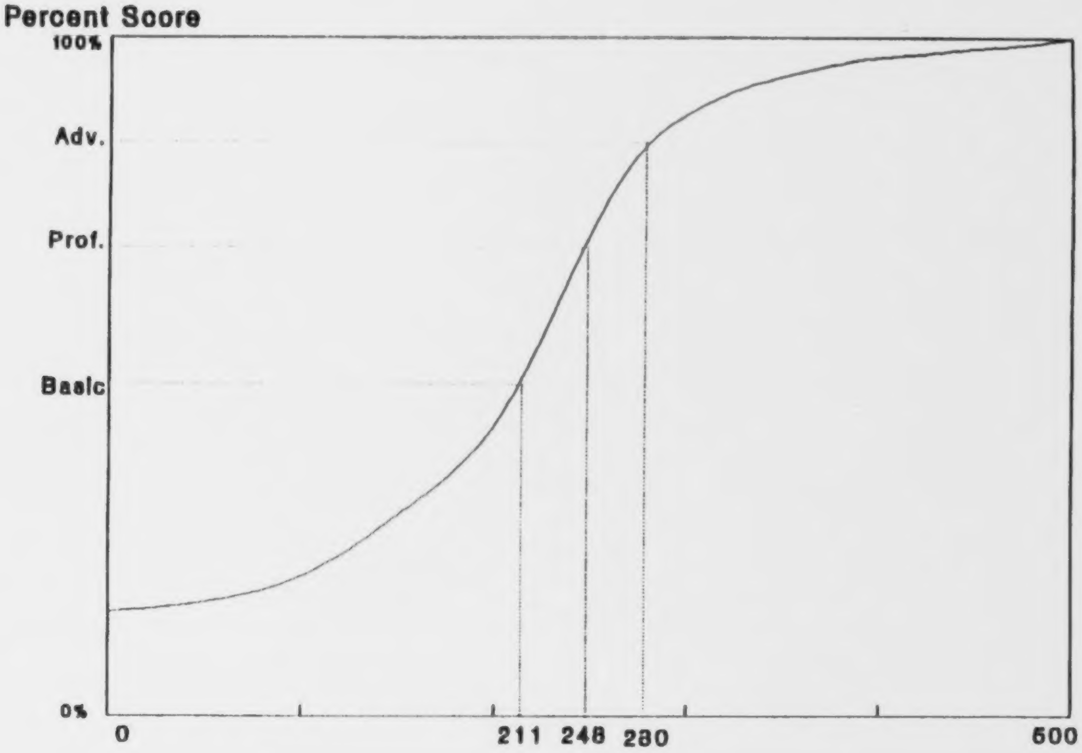


Figure 3. Using performance standards to increase the meaningfulness of NAEP scores.

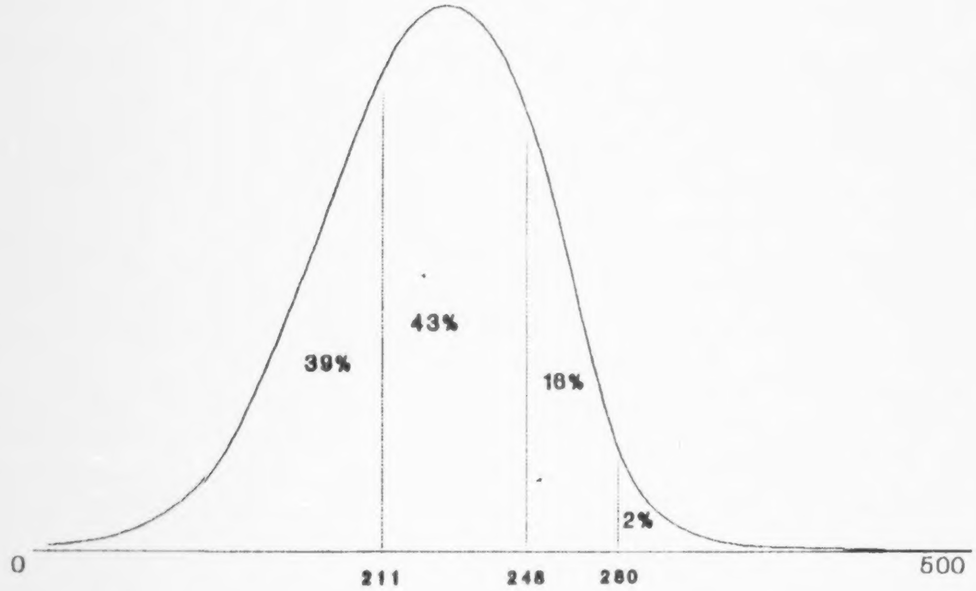


Figure 4. (Approximate) distribution of 1992 grade 4 NAEP mathematics results.

reported and used in ways that interested audiences will understand and interpret correctly (see Wainer,1992, for examples of problems in reporting data).

Our research described in this portion of the paper was funded by the National Center for Education Statistics (NCES) and was stimulated by several recent studies of NAEP reports which found that policymakers and the media were misinterpreting some of the texts, figures, and tables (Jaeger, 1992; Linn & Dunbar, 1992; Koretz & Deibert, 1993). Our purposes in this study were (a) to investigate the extent to which NAEP executive summaries were understandable to policymakers, educators, and the media, and the extent to which problems were identified; and (b) to offer a set of recommendations for improving performance-standard-based reporting practices. Such a study seemed essential because there is an unevenness in the measurement literature: Relatively large amounts of literature exist on a variety of technical topics such as test development, reliability, validity, standard setting, and proficiency estimation, but relatively little work has been done on the topic of reporting test score information to communicate effectively with a variety of audiences (for an exception, see Aschbacher & Herman, 1991). More research is needed to provide a basis for the development of guidelines. This study was a modest first step toward the goal of improving test score reporting.

Basic Methodology

The interview used in this study was designed around the *Executive Summary of the NAEP 1992 Mathematics Report Card for the Nation and the States* (Mullis et al., 1993). This report was chosen because it was relatively brief and could stand alone for policymakers and educators. In addition, the NAEP Executive Summary reports are well known and widely distributed (over 100,000 copies of each Executive Summary are produced) to many people involved in various areas of education. Further, we thought that the NAEP Executive Summary results that included both national and state results would be of interest to the interviewees, who were from different areas of the country. Like most executive summaries, this report's format contains tables, charts, and text to present only the major findings of the assessment. For a more in-depth analysis of the NAEP 1992 mathematics results, readers need to refer to some of the more comprehensive NAEP reports such as National Center for Education Statistics (1993).

Our goal in the interviews was to determine just how much of the information reported in the Executive Summary was understandable to the intended audience. We wanted to attempt to pinpoint the aspects of reporting that might have been confusing to the readers and to identify changes in the reporting that the interviewees would like to see.

The 1992 NAEP mathematics Executive Summary report consists of six sections that highlight the findings from different aspects of the assessment. We designed interview questions for each section in an attempt to ascertain the kind of information interviewees were obtaining from the report. We asked interviewees to read a brief section of the report, and then we questioned them on the general meaning of the text or on the specific meaning of certain phrases. Interviewees also examined tables and charts and were asked to interpret some of the numbers and symbols. Throughout the interviews, we encouraged the interviewees to volunteer their opinions or suggestions. This kind of information helped us gain a general sense of what the interviewees felt was helpful or harmful to them when trying to understand statistical information.

The 60 participants in the interviews represented a broad audience similar to the intended audience of the NAEP Executive Summary reports. We interviewed policymakers, educators, and people in the media in Massachusetts; Connecticut; Washington, D.C.; Louisiana, Kentucky; and New York. We spoke with people at state departments of education, attorneys, directors of companies, state politicians and legislative assistants, school superintendents, education reporters, and directors of public relations. Many of the people we interviewed were prominent individuals in their fields, and most held advanced degrees. Despite this, however, many interviewees had problems reading and interpreting the information they were shown.

Major Findings

The interviewees in this study seemed very interested and willing to participate. Most of them regularly received reports like the Executive Summary in their offices. They were eager to help us to determine the extent to which these reports were understandable and to be involved in the improvement of these reports by offering their opinions.

We found that, despite the fact that the interviewees tried to understand the report, many of them made fundamental mistakes. Nearly all were able to generally understand the text in the report, though many would have liked to have seen more descriptive information (e.g., definitions of measurement and statistical jargon and concrete examples). The problems in understanding the text involved the use of statistical jargon in the report. This confused and even intimidated many of the interviewees. Some mentioned that, although they realized that certain terms were important to statisticians, those terms were meaningless to them. After years of seeing those terms in reports, they tended to "glaze over" them.

The tables were more problematic than the text for most of the interviewees. Although most got a general feeling of what the data in the tables meant, the interviewees made many mistakes when we asked them specific questions. The symbols in the tables (e.g., to denote statistical significance) confused some; others just chose to disregard them. For example, interviewees often "eyeballed" the numbers to determine if there was improvement, ignoring the symbols next to the numbers denotating statistical significance. Improvement to these interviewees often meant a numerical increase of any magnitude from one year to the next.

Consider Table 1 from the Executive Summary, which is reproduced as Table 2 in this paper. Policymakers, educators, and the media alike indicated several sources of confusion:

1. They were baffled by the reporting of average proficiency scores (few understood the 500-point scale). Also, interviewees confused proficiency as measured by NAEP and reported on the NAEP scale with the category of "proficient students."

2. Interviewees were baffled by the standard error beside each percentage. These were confusing because (a) they got in the way of reading the percentages, and (b) the notes did not clearly explain to the interviewees what a standard error is and how it can be used.

Table 2

National Overall Average Mathematics Proficiency and Achievement Levels, Grades 4, 8, and 12

Grade	Assessment Year	Average Proficiency	Percentage of Students at or Above			Percentage Below Basic
			Advanced	Proficient	Basic	
4	1992	218(0.7) >	2(0.3)	18(1.0) >	61(1.0) >	39(1.0) <
	1990	213(0.9)	1(0.4)	13(1.1)	54(1.4)	46(1.4)
8	1992	268(0.9) >	4(0.4)	25(1.0) >	63(1.1) >	37(1.1) <
	1990	263(1.3)	2(0.4)	20(1.1)	58(1.4)	42(1.4)
12	1992	299(0.9) >	2(0.3)	16(0.9)	64(1.2) >	36(1.2) <
	1990	294(1.1)	2(0.3)	13(1.0)	59(.5)	41(1.5)

> The value for 1992 was significantly higher than the value for 1990 at about the 95 percent confidence level.

< The value for 1992 was significantly lower than the value for 1990 at about the 95 percent confidence level. The standard errors of the estimated percentages and proficiencies appear in parentheses. It can be said with 95 percent confidence that for each population of interest, the value for the whole population is within plus or minus two standard errors of the estimate for the sample. In comparing two estimates, one must use the standard error of the difference.

3. Most interviewees misunderstood or ignored the < and > signs. Even after reading the notes, many interviewees indicated that they were still unclear about the meaning.

4. Interviewees were most confused about the reporting of students *at or above* each proficiency category. They interpreted these cumulative percents as the percent of students in *each* proficiency category. Then they were surprised and confused when the sum of percents across any row in Table 2 did not equal 100%. Contributing to the confusion in Table 2 was the presentation of the categories in the reverse order to what was expected (i.e., Below Basic, Basic, Proficient, and Advanced). This information, as presented, required reading from right to left instead of from the more common left to right. Perhaps only about 10% of the interviewees were able to make the correct interpretations of the percents in the table.

5. Interviewees did not always read notes, and they often misunderstood the notes when they did read them.

6. Some interviewees expressed confusion because of variations between the NAEP reports and their own state reports.

We prepared Table 3 to respond to many of the criticisms the interviewees raised about Table 2. Modest field testing during the study indicated that Table 3 was considerably less confusing. A simplified Table 3 may be more useful to intended audiences of the report, but Table 3 may be inconsistent with the reporting requirements of a statistical agency such as NCES.

Another common problem for the interviewees was reading the charts. In an assessment of national scope, it is often necessary to include quite a bit of information in each chart. This requires the use of some elegant graphic techniques. It also tends to add to the complexity of the charts. Although these charts are impressive in the NAEP report, they were intimidating to those who could not interpret them. The unfamiliar chart formats were very difficult for many of the interviewees. Once the charts were explained to them, the interviewees understood them, but many commented either that they couldn't have figured the charts out on their own or, more commonly, that they simply would not have the time in a typical day to devote to a report requiring so much study.

Here again, the notes were of little help in explaining the tables and charts. They were often lengthy and contained statistical explanations that the interviewees did not understand. For example, many of the interviewees found the following note particularly confusing:

> The between state comparisons take into account sampling and measurement error and that each state is being compared with every other state. Significance is determined by an application of the Bonferroni procedure based on 946 comparisons by comparing the difference between the two means with four times the square root of the sum of the squared standard errors. (Mullis et al., 1993, Fig. 1, p. 12)

The first sentence of this note would have been sufficient for the audience we interviewed.

Table 3

National Overall Average Mathematics Proficiency and Achievement Levels, Grades 4, 8, and 12

Grade	Assessment Year	Average Proficiency	Percentage of Students at or Above			
			Below Basic	Basic	Proficient	Advanced
4	1992	218 >	39%	43%	16%	2%
	1990	213	46	41	12	1
8	1992	268 >	37%	38%	21%	4%
	1990	263)	42	38	18	2
12	1992	299(0.9) >	36%	48%	14%	2%
	1990	294(1.1)	41	46	11	2

The symbols ">" and "<" are used to highlight differences in the table that are large enough to be real and not due to chance factors such as instability in the information. For example, it can be said that average mathematics performance in Grade 4 in 1992 was higher than in 1990.

Despite the fact that many of the interviewees made mistakes, their overall reactions to the task were positive. Some were surprised to find that when they took the time to look at the report closely, they could understand more than they expected. Again, however, most noted that they did not have the time needed to scrutinize the reports until they could understand them. When we apologized to one legislator for the short time we may have allowed for the task, he noted that he had already spent more time with us than he would have spent on his own with the report.

The interviewees who had problems found the results easier to understand once we explained some of the tables and statistical concepts to them. Of course, there were a few interviewees who became so frustrated with the report or with themselves that they simply gave up trying to understand it.

Everyone offered helpful and insightful opinions about the report. Some common suggestions emerged about how to make the results in reports like the Executive Summary more accessible to those with little statistical background. A comment made by a couple of interviewees was that the report appeared to be "written by statisticians, for statisticians." As a remedy, many suggested removing the statistical jargon. It seems that phrases such as "statistically significant" do not mean much to the audience we interviewed and often only intimidated the readers.

Another suggestion was to simplify the tables by placing the standard errors in an appendix. The lengthy notes could also be placed in an appendix, as these tended to clutter the appearance of tables. Many interviewees would prefer brief notes in layperson's terms. Also, according to many interviewees, it would be better to present some of the information in simple graphs instead of in tables. One reason is that a simple graph can be understood relatively quickly.

From some of the comments mentioned above, it is apparent that most interviewees needed to be able to quickly and easily understand reports. They simply did not have much time or were unwilling to spend much time. Some interviewees would even prefer receiving a more lengthy report if it were just a bit clearer and easier to understand.

Among our tentative conclusions from the study are that (a) there was a considerable amount of misunderstanding about standards-based reporting in the NAEP mathematics assessment Executive Summary we studied; (b) improvements will need to include the preparation of substantially more user-friendly reports with considerably simplified figures and tables; and (c) regardless of the technical skills of the audiences, reports ought to be kept straightforward, short, and clear because of the short time people are likely to spend with these executive summaries.

On the basis of our limited and preliminary research, we offer several reporting guidelines for NAEP and state assessments:

1. Charts, figures, and tables should be understandable without having to refer to the text. (Readers didn't seem willing to search around the text for interpretations.)

2. Always field-test graphs, figures, and tables on focus groups representing the intended audiences; much can be learned from field-testing report forms. (The situation is analogous to field-testing assessment materials prior to their use. No respectable testing

agency would ever administer important tests without first field-testing their material. The same guideline should hold for the design of report forms.)

3. Be sure that charts, figures, and tables can be reproduced and reduced without loss of quality. (This is important because interesting results will be copied and distributed and no one wants to look at bad copies. Correct interpretations, let alone interest, can hardly be expected if copy quality is bad.)

4. Graphs, figures, and tables should be kept relatively simple and straightforward to minimize confusion and shorten the time readers need to identify the main trends in the data.

Currently, we are preparing a final report of our research in which more details on the research study will be provided, along with an expanded set of score-reporting guidelines (Hambleton & Slater, 1995).

CONCLUSIONS

In principle, standards-based reporting provides policymakers, educators, the media, and the public with valuable information. But the burden is on the reporting agency to ensure that the reporting scales used are meaningful to the intended audiences and that the reported scores are valid for the recommended uses. At the same time, reporting agencies need to focus considerable attention on the way in which scores are reported to minimize confusion as well as misinterpretations and to maximize the likelihood that the intended interpretations are made. Ensuring that the reports are interpreted fully and correctly will require the adoption and implementation of a set of guidelines for standards-based reporting that include field-testing of all reports. Special attention will need to be given to the use of figures and tables, which can convey substantial amounts of data clearly if they are properly designed. *Properly designed* means that the figures and tables are clear to the audiences for whom they are intended.

The *Adult Literacy in America Study* (Kirsch, Jungeblut, Jenkins, & Kolstad, 1993), conducted by NCES, Westat, and Educational Testing Service, appears to have benefited from some of the earlier evaluations of NAEP reporting and provides some excellent examples of data reporting. A broad program of research involving measurement specialists, graphic design specialists (see, for example, Cleveland, 1985), and focus groups representing intended audiences for reports is very much in order to build on some of the successes in reporting represented in the *Adult Literacy in America Study* and some of the useful findings reported by Jaeger (1992), Koretz and Deibert (1993), and others. Ways need to be found to balance statistical rigor and accuracy in reporting with the informational needs, time constraints, and quantitative literacy of intended audiences.

These are potentially important times for measurement specialists. Policymakers and the public seem genuinely interested in assessment results. But without improvements to the scales and reporting forms employed, no matter how well tests are constructed and data are analyzed, there is the serious risk that assessment results will be ignored, misunderstood, or judged irrelevant. The challenge to measurement specialists is clear. It is now necessary to proceed with the essential research.

References

American College Testing. (1993). *Setting achievement levels on the 1992 National Assessment of Educational Progress in mathematics, reading, and writing: A technical report on reliability and validity*. Iowa City: Author.

Aschbacher, P. R., & Herman, J. L. (1991). *Guidelines for effective score reporting* (CSE Tech. Rep. No. 326), Los Angeles: UCLA Center for Research on Evaluation, Standards, and Student Testing.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17*(12), 1991-204.

Beaton, A. E., & Johnson, E. G. (1993). Overview of the scaling methodology used in the National Assessment. *Journal of Educational Measurement, 29*(12), 163-176.

Bourque, M. L., & Garrison, H. H. (1991). *The levels of mathematics achievement: Vol. 1. National and state summaries*. Washington, DC: National Assessment Governing Board.

Cleveland, W. S. (1985). *The elements of graphing data*. Monterey, CA: Wadsworth.

Forsyth, R. A. (1991). Do NAEP scales yield valid criterion-referenced interpretations? *Educational Measurement: Issues and Practice, 10*(3), 3-9, 16.

Hambleton, R. K. (1994, April). *Scales, scores, and reporting forms to enhance the utility of educational testing*. Paper presented at the annual meeting of National Council on Measurement in Education, New Orleans.

Hambleton, R. K., & Bourque, M. L. (1991). *The levels of mathematics achievement: Initial performance standards for the 1990 NAEP Mathematics Assessment*. Washington, DC: National Assessment Governing Board.

Hambleton, R. K., & Slater, S. (1995). *Reporting NAEP results to policy makers and educators: Is the information understood*? (Laboratory of Psychometric and Evaluative Res. Rep. No. 271). Amherst, MA: University of Massachusetts, School of Education.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamental of item response theory*. Newbury Park, CA: Sage.

Jaeger, R. M. (1992). General issues in reporting of the NAEP trial state assessment results, in R. Glaser & R. L. Linn (Eds.), *Assessing student achievement in the states* (pp. 107-109). Stanford, CA: National Academy of Education.

Kirsch, I. S., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). *Adult literacy in America: A first look at the results of the National Adult Literacy Survey*. Washington, DC: U.S. Government Printing Office.

Koretz, D., & Deibert, E. (1993). *Interpretations of National Assessment of Educational Progress (NAEP) anchor points and achievement levels by the print media in 1991*. Santa Monica, CA: Rand.

Linn, R. L., & Dunbar, S. B. (1992). Issues in the design and reporting of the National Assessment of Educational Progress. *Journal of Educational Measurement, 29*(2), 177-194.

Mullis, I. V. S. (1991, April). *The NAEP scale anchoring process for the 1990 Mathematics Assessment*. Paper presented at the meeting of the American Educational Research Association, Chicago.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., & Phillips, G. W. (1993). NAEP 1992 mathematics report card for the nation and the states. Washington, DC: National Center for Education Statistics.

National Center for Education Statistics. (1993). *NAEP 1992 mathematics report card for the nation and the states* (Rep. No. 23-ST02). Washington, DC: Author.

Phillips, G. W., Mullis, I. V. S., Bourque, M. L., Williams, P. L., Hambleton, R. K., Owen, E. H., & Barton, P. E. (1993). *Interpreting NAEP scales*. Washington, DC: U.S. Department of Education.

Sufflebeam, D. L., Jaeger, R. M., & Scriven, M. (1991). *Summative evaluation of the National Assessment Governing Board's inaugural 1990-91 effort to set achievement levels on the National Assessment of Educational Progress*. Kalamazoo, MI: Western Michigan University.

Wainer, H. (1992). Understanding graphs and tables. *Educational Researcher, 21*(1), 14-23.

Using Performance Standards to Report National and State Assessment Data: Are the Reports Understandable and How Can They Be Improved?

Summary of Break-out Session[1]

The group began its discussion on *ways to improve the reporting* of national and state assessment data. One participant observed that there is a need for more information rather than a need for simplifying information. She indicated that there is still a dichotomy between having a lot of information and consuming a little. Hambleton suggested that it is important to provide consumers with meaningful information presented in a simple form. He further suggested that, given the purposes of the National Assessment of Educational Progress (NAEP), states should start linking their assessments to the NAEP. Some participants argued that either only bottom-line cut points and not error bands be reported and/or that the score scale be changed. Others countered that these data give little information to the public.

Hambleton stated that the National Center for Education Statistics (NCES) has a role dilemma: reporting data to the public versus being a statistical agency. The participants stated that there is also a conflict in what the public wants to know and what NAEP purports to do. It was suggested that NCES should provide some additional interpretative information related to NAEP data for its different audiences.

The discussion addressed ways to simplify NAEP reporting. It was pointed out that NAEP releases, both at the national and state levels, could provide short synopses to capsulize findings; these reports should use the mean score. Other participants suggested that in the reports, ranking charts could be provided as incentives to the states. The group generated additional suggestions as follows: (a) the use of simple charts with sentences on how to read the tables, (b) the use of achievement levels instead of anchor points (the latter are too technical), and (c) development of a strategy for teaching the public and the media how to understand the basics of measurement (e.g., the meaning of proficiency levels). In addition, the group agreed that the targeted audience of the NAEP reports should be made explicit.

The group discussed issues related to the validation of the results. Specific reference was made to the 1988 release of NAEP data by the Secretary of Education and of an interpretation of the performance of American students in relation to students in other countries. Hambleton explained that mathematics achievement scores were released and then equating studies were done that generated the opposite conclusions. The discussion raised the question of whether the standards are too high or unreasonable. The discussion focused on the difficulties and complexities of equating (using different methods).

The continuing discussion focused on the professional responsibilities of federal statistical agencies. It was suggested that historically, statistical agencies have been relied upon to produce a range of data. Hambleton indicated that NCES has standards for every report. It was suggested that these agencies continue to redefine themselves as the state of the art of testing and assessment changes, and as the debate over NAEP continues. Providers of information bear the responsibility to make it useful.

---

[1] This is a summary of the break-out discussions for Ronald Hambleton's presentation. The session was facilitated by Susan Ahmed (National Center for Education Statistics) and recorded by Kristen Keough (National Center for Education Statistics).

The group suggested that there is a need for more research, including focus groups, to help NCES plan. A representative from NCES pointed out that the next proposed wave of activities will include focus groups to make research more accessible through short publications, such as "NAEP Facts" for teachers. However, he indicated that the continuing questions for NCES are: Who comprises the audience that we serve, and can we serve the audience efficiently?

The Impact of Standards on How Teachers Behave in the Classroom: Promises and Perils

"For every complex issue there is a simple answer, and it is wrong." (McMillan, 1994, p. 466)

Phyllis W. Aldrich

Coordinator of the Gifted and Talented Program, Director of the National Javits Language Arts Research Project, WSWHE Board of Cooperative Educational Services

ABSTRACT

The delineation of standards at a national level could have a powerful and positive effect on what actually happens in the classroom, if teachers are directly involved at their schools in discussing interpretation and potential applications to their own specific students and curriculum. Expectations for student achievement and standards are always present but rarely articulated and not readily comparable. Careful description of student performance levels, clearly described with many examples, could provide valuable "calibration" or "anchors." This would inform individual teachers on what might be possible for students to achieve at a specific grade level in a specific area of study. The linkage between nationally derived student performance levels and adaptation to local practices could be a useful tool in the campaign to improve student learning.

The central focus of this paper is an analysis of the question: "Could student performance standards lead the way to improved curriculum and instruction?"

SARAH'S QUESTION

After spending five months in a ninth grade far from home, Sarah returned to her local high school to study the same subjects, which included French. After the first week of school, she came home indignantly sputtering about a test in her French class. "My friends here are being gypped. We all got an *A* on the test yesterday, and they think they are learning French well, and they are proud of the *A*. But answers like ours on a French test at my other school would barely earn a *C*! That teacher wouldn't let us get away with such a lame performance. She wouldn't have asked such easy questions. Sure, the kids may complain about a lot of homework, but right now they are being fooled about what they are learning and it's a gyp. If no one tells them, they will still think they are doing OK and they are NOT! Is this fair?"

CEDRIC'S STRUGGLE

A stark description of the drastic consequences of the uneven standards troubling Sarah was highlighted in a recent front-page story in *The Wall Street Journal* (Suskind, 1994) about 17-year-old Cedric Jennings. Cedric is a black student from a Washington, DC, high school who had a wrenching experience at Massachusetts Institute of Technology's (MIT) summer program for promising minority students. Cedric had aspired to MIT for years as a path out of the ghetto but found during the summer of intense studies designed to strengthen his skills that he was still way behind his peers. At the end of the program, he was told he could not measure up to MIT's admissions standards. The reporter noted that "despite years of asking for extra work after school--of creating his own independent-study course just to get the basic education that other students elsewhere take for granted--he [Cedric] was woefully way behind" (p.1). In a curriculum and school culture that does not uphold quality standards, even heroic efforts by hardworking students with supportive parents are not enough.

To understand the extent of the nation's educational failings, reformers should go beyond generalities of economic competitiveness to remember that the lack of common high standards for teaching and learning has tragic consequences for students. But, will the definition of student achievement standards lead us to high-quality learning and instruction? Or as Cuban (1993) claims, is it a mistake to assume that "curricular change [could] transform teacher behavior and student learning" (p. 183)?

STANDARDS AS HARMFUL FANTASY OR NORTH STAR?

Does the focus of some reformers on standards and curriculum reform constitute a "harmful fantasy" (Cuban, 1993, p. 184), a distraction from "the deeper analysis of school reform" (Eisner, 1994, p. 15), or grounds for "widespread consternation" (Standards and assessment, 1993, p. 6) as some critics have suggested? Or could the clear delineation of expected levels of student performance in each discipline provide an indispensable North Star to keep reform efforts on course (Simmons & Resnick, 1993)? A central issue seems to be whether curriculum and assessment reform efforts will make any difference in what actually happens to students. Some (Cuban, 1993) claim that a concentration on achievement standards shunts policymakers to the peripheral role of those who "fiddle with official curricula because these are all they can really influence" (p. 184). Such scholars doubt that the vision of higher standards can be easily translated from seminars and conferences into teachers' heads and hearts.

Cuban (1993) suggests that if reformers deliberate only on issues of the "official" curriculum and ignore the power of three other forms of curricula, nothing will change. His contention is that in addition to the officially defined curriculum, there are three other powerful forms: those that are actually "taught, learned and tested" (p. 184). While all four ways of looking at curricula are often contradictory, all need to be understood and aligned if meaningful change is to occur. Darling-Hammond (1991) reminds us that any analysis of policy about setting standards cannot afford to overlook considerations of the personal qualities of teachers, the art of teaching and the idiosyncrasies of individual students. If the reform conversation concentrates primarily on either the technical complexities of testing methodology or the content of curriculum frameworks, Cuban warns that reformers may be doomed to be remembered as one more passing phase in the "pitiful history" (p. 182) of curricular reform.

WISELY DEVELOPED STANDARDS MIGHT HELP

Mirel and Angus suggest a more optimistic view about the role of national goals and standards in effecting change. Their 1994 study of American high schools over the last 100 years concluded that "national goals and standards, wisely developed and applied, can greatly benefit American education" (p. 4). They looked closely at the issue of equalizing educational quality for all students and argued that rather than shutting out poor and minority students, "clearly articulated national content and performance standards and well-designed national methods of assessment" (p. 184) can (and have) enhanced opportunity for all. They disagree with Cuban's (1993) contention that "heated policy debates over national curriculum" (p. 184) diverts attention from equity issues. Rather, they would concur with proponents of standards reform and equity such as Warren Simmons from the New Standards Project.

In a "conversation" with John O'Neil from *Educational Leadership* (1993), Simmons argues that rather than hurting minority students, standards are helpful if they are not held to high standards by employers, by their communities, and so forth. So the idea that the educators are doing students a disservice by holding them to high standards is a fallacy" (O'Neil, p. 20).

In their recommendations about the importance of designing ways to translate rich and rigorous curricular standards into classrooms, writers such as Mirel and Angus (1994) and Simmons and Resnick (1993) share some common ground with Cuban (1993). They all agree that both considerations of national standards and pedagogy are important. It is my position that standards can have a significant impact but cannot be grafted onto fiercely guarded local prerogatives. Standards need to become an intrinsic part of the school culture and teacher belief systems. For that to happen, policymakers need a fuller understanding of the complex layers of curricula and how climates for change can be facilitated.

TRANSLATING STANDARDS INTO CLASSROOM PRACTICE

The challenge for school reformers today is indeed complex. Determining where to concentrate efforts is a baffling process. Mitchell (1992) has highlighted the scope of the problem by reminding us that there are about 2.3 million teachers in the United States, most of whom will still be teaching 10 years from now. To borrow Deborah Meier's phrase, from a speech in 1994, we are faced with the challenge of figuring out "how to change the tire while the car is moving." Teacher beliefs, their "personal theorizing," skills, knowledge, and experience will continue to be powerful forces affecting how the official curriculum is filtered into actual lessons and learning (Ross, 1992).

If we are to move forward to raise student achievement through national standard-setting initiatives, we will need simultaneously to develop strategies to influence the entire school culture. Mitchell (1992) warns that while clear definition of world-class standards is a vital component of assessment and a necessary factor in school reform, it is not sufficient.

If the entire system is to be changed, the campaign needs to be waged on many different battlefields at once: national, state, and local. Reformers will need to pursue credible, high national standards for student achievement while investigating creative and practical ways for teachers to use them. To succeed, both the presence of rich content frameworks, derived through the national professional and scholarly consensus processes, and their thoughtful translation through state and local leadership into teachers' daily instruction are needed. How can this task be accomplished?

CHANGING A SCHOOL CULTURE

Resnick's (1992) intensive pioneering efforts in the past four years to build a successful New Standards Project with teachers from 18 states and 5 districts are based on an optimistic premise:

> If we can agree on national standards for student achievement and create conditions in school systems all over the country in which those standards are internalized and made the centerpiece of educators' and students' efforts, there is a good probability that curriculum, professional development, textbooks, and, eventually teacher preparation can be changed so that the entire system is working toward the standards. (p. 412)

The major challenges are at least twofold: how to establish common standards for students, and how to make them central to the life of a classroom. For some schools, there is also the issue of how to move beyond the official standards.

In order to move toward the goal of dramatically raising student achievement, many policy leaders and scholars have stressed the need to set clear targets (Finn & Ravitch, 1994). For example, as Colorado Governor Romer (1994) has so vigorously noted, both content standards *and* performance standards are needed. Reformers need to know "how good is good enough" (p. 1). Mitchell (1992) believes that both vision from policy leaders as to the importance of national, world-class standards and concrete definitions of what those standards might look like will be critical.

THE ROLE OF A NATIONAL ASSESSMENT SYSTEM

Historically, American schools have shunned the concept of nationally imposed standards for student performance as an unwelcome form of federal control. Determination of student achievement has been left to a patchwork system of idiosyncratic teacher grading practices with criteria varying from grade to grade and school to school (Crooks, 1988; Office of Educational Research and Improvement, 1994; Spillane, 1993). Often the teachers' manuals accompanying textbooks have effected a *de facto* standardized set of student performance expectations, but the same texts are not used in all schools, nor has there been any agreement among scholars or teachers that the texts cover all the important competencies.

Norm-based achievement tests, such as the Iowa Test of Basic Skills or the California Achievement Test, while appearing to yield a standardized measure of how well students know their math and reading, have, over the years, led school administrators to draw the incongruous conclusions about students performance chronicled by Cannell in 1989 as the "Lake Woebegone Effect." College admissions tests, such as the Scholastic Aptitude Testing, (SAT) and American College Test (ACT), while standardized nationally, are taken only by college-bound students and do not actually measure the degree of mastery of "substantive subject matter" (Cheney, n.d., p. 6). In contrast, the national school-leaving examinations taken by high school students in Europe and Asia require in-depth knowledge of specific subjects (Cheney, 1991). Over the years, American schools have relied primarily on teachers' subjective expectations for student performance. These expectations have not been linked to any commonly articulated frame of reference, have not been readily comparable between schools or states, and have rarely been shared with students or the public (Cizek, 1993).

This lack of linkage was an important factor in the 1969 legislation to create the National Assessment of Education Progress (NAEP). For the past 22 years, NAEP has monitored on a national basis student knowledge and skills through biennial sampling of students in grades 4, 8, and 12 in the basic subjects of reading, writing, math, science, and history (Mullis, Owen, & Phillips, 1990).

A Nation at Risk (Gardner, 1983) warned that student achievement was still mediocre. In 1988 Congress passed legislation that created the National Assessment Governing Board (NAGB) to develop policy for the expansion of NAEP, to devise assessment frameworks through a national consensus process, to produce state-by-state comparison data, and to explore ways to determine achievement levels. The NAGB mission was not only to measure what students knew and were able to do, but also to define ways to judge whether the measured achievement was good enough (Owens, 1988). In 1990, NAGB initiated trial analyses of student achievement levels in mathematics at grade 8 (Bourque &

Garrison, 1991). In 1992, analysis of student achievement was extended to include math at grades 4 and 8 and reading at grade 4.

Reports from the first two studies on achievement levels occurred amid protracted and noisy debate about how, or indeed whether, achievement levels should or could be defined by any federal agency. Along with issues of technical validity of the standard-setting process, discussion about whether states or local schools would or should be guided by nationally derived achievement levels continues to be lively and complex.

Proponents for high national standards for student achievement argue for the importance of a "top-down" determination of achievement levels to fill the void created by the lack of interstate and interteacher agreement. Opponents, who may agree on the need for new assessment approaches, are convinced that a nationally imposed system cannot be grafted onto fiercely guarded local prerogatives. They argue persuasively that teachers must be directly involved at their local school level in discussing interpretation and potential applications to their own specific students and curricula (O'Neil, 1994). Is it possible as Redfield (1989) has suggested that "top-down" and "bottom-up" approaches might work together?

HOW MIGHT PERFORMANCE STANDARDS AFFECT CURRICULUM POLICY?

Another general question addressed by the standard-setting conference was, "What effects might large-scale student performance standards have on issues of curriculum policy?" Even if reformers successfully surmount the hurdles of resolving the complex, technical issues in large-scale assessments and arrive at solid standards, and assuming that understanding is established among psychometricians and the public, will instruction improve? Will students learn more? Will schools be able to document greater student achievement? Maybe.

ASSUMPTIONS ABOUT CURRICULUM

A particular focus for this paper is how standards might impact teacher decisions about *what* to teach and *how* to teach actual students when the classroom door has closed after the speeches and workshops have ended. Tyler's 1949 classic study of curriculum defined it as the set of intended learning outcomes or instructional objectives of a program. More recently, the New York State Board of Regents (1994) has framed a far more comprehensive definition of curriculum: "All the arrangements a school makes for students' learning and development, including the sequence, format, and content of courses, student activities, teaching approaches, books, materials, and resources used, and the way in which teachers and classes are organized, which enable students to reach standards" (p. 1). With this definition, it is obvious that the issues surrounding achievement standards pose profound challenges for supervisors, curriculum directors, and principals as well as teachers.

Darling-Hammond (1992) has acknowledged the importance of standards but proposes that their effectiveness is derived from perceptions about "who controls assessment" (p. 26). Externally imposed mandates won't work. She predicts that even "more challenging and thought-provoking performance-based assessments will fail to transform schools if assessments are externally mandated and delivered" (p. 26). She sees the creation of national, curriculum-standard frameworks only as a first step. While the delineation of standards at a national level could have a powerful and positive effect on what actually happens in the classroom, she predicts that "genuine accountability" will only be realized when schools reorganize themselves as inquiring, collaborative "communities of democratic discourse" (p. 27).

SCHOOLS AS INQUIRING COMMUNITIES FOR TEACHERS AND STUDENTS

How might that community be created? One way is for teachers to analyze student work to determine benchmarks along a continuum of mastery. When making initial decisions about what and how to teach in a specific subject, teachers need to probe issues of "how good is good enough" (Romer, 1994, p. 1). The New Standards Project is predicated upon the assumption that samples of student work at different achievement levels can serve as valuable intra- and interstate and regional benchmarks.

NAGB approaches the issue by suggesting that careful description and clear exemplars of student performance levels set by nationally convened consensus groups could provide valuable "calibration" or "anchors" to inform individual teacher efforts to know what might be possible for students to achieve at a specific grade level in a specific area of study. A possible extension of the NAGB achievement levels work is to help teachers satisfy their natural curiosity about what occurs in classrooms beyond their own. Why one teacher awards an "A" to a paper may not be readily obvious to another. Without commonly described achievement levels, too many of the factors that underlie teacher judgments remain unarticulated, not only among each other, but also among their students. Hence real mastery is hard to determine.

One of the benefits reported by teachers who participate in scoring large-scale student assessments such as NAEP, the College Board Advanced Placement Examination, or the Vermont writing or math portfolio projects, is that the criteria for making determinations between papers rated as a "3" or "5" or between a "minimal" or "elegant" math solution are discussed. Revisiting ratings during the scoring process helps teachers clarify their own thinking (Hewitt, 1994). The scoring rubrics for the NAEP 1990 writing portfolio have attracted teacher interest as a guide to share with students before they venture into any writing assignment (Gentile, 1994). Some teachers described in *The New Standards Project* (1994) have already found that the availability of sample student responses deemed "minimal," "proficient," or "advanced" by colleagues in other schools are useful guides to teachers as they formulate expectations for student performance in their own classes.

The issue of improving student achievement should be approached as a central part of the larger school culture. There are many curriculum-related factors that must be considered in order to change teacher beliefs and behaviors. Commonly understood or "moderated" assessment standards will not be sufficient by themselves, but they will be vitally important because without a clear destination and a North Star to navigate by, reformers will not know where they are or whether they have arrived (Mitchell, 1992).

What if voluntary calibration opportunities on a national level became practical and economical? What if the generation of banks of new assessment tasks in projects such as New Standards, which involves interstate teacher cooperation, continued to thrive (Flanagan, 1994)? Finally, what if more opportunities were created for state-level leaders to engage with national frameworks in the kind of collaboration envisioned by a recent initiative of the Council of Chief State School Officers? Then what?

TEACHER ENGAGEMENT WITH FRAMEWORKS

Even if these ideal conditions come to pass, the local curriculum must necessarily be targeted for real change to occur (Jamentz, 1994). Intensive and ongoing commitment to staff development, with time for monitored experimentation with the new assessment tools, time to practice, to talk to colleagues, to reflect, to "internalize" new ways to conceive of gathering evidence about student achievement will be necessary (Resnick, 1992).

The New Hampshire Education Improvement and Assessment Program provides a good example of how a statewide educational accountability system relies on curriculum frameworks to effect improvements in the learning of all students. In her 1994 report to the New Hampshire Program Improvement Partnership, Lachat describes the projected four curriculum frameworks that "include broad goals, content standards which define what children should be able to do, and identify proficiency standards to be measured at the end of third, sixth, and tenth grades" (p. 2). The curriculum frameworks are intended to raise expectations for all students by describing in detail the important ideas and skills in each subject. Lachat predicts that schools no longer will have to depend on "textbook publishers to define the scope and sequence of what is taught" (p. 5).

New Hampshire has developed an assessment system that is aligned with the standards spelled out in the curriculum frameworks. Once this system is fully implemented, student performance will be compared to fixed standards of achievement that have four proficiency levels: advanced, proficient, basic, and novice. Districts will receive individual student reports to share with parents and to identify where a district needs to improve its instructional program. The state guidelines stress that since the 1994 assessment for grades K-3, will "represent new and higher standards for student learning" (Lachat, 1994, p. 10), many students who used to test well on earlier norm-referenced tests may score mainly at lower levels.

Because this is the first time students have been measured against new rigorous standards, districts are being urged to perceive the results as a source of "important base-line data" (Lachat, 1994, p. 10). The New Hampshire results are intended to pinpoint district needs for staff development and curriculum improvement rather than serve as a trigger for public alarm. The public reception of this statewide adoption of a coherent approach to the challenge of using large-scale assessments to improve curricula will be important to watch in the coming months.

An essential ingredient in this reform process will be vigorous teacher discussion of elements of various state and national curriculum frameworks. It will be important for teachers to see the entire spectrum of the curriculum K-12 in order to better understand where their decisions about what to teach fit into the larger picture. New York recently released a single math-science-technology curriculum framework that describes all three fields in an integrated way so that students' learning can occur in a related context. If school administrators approach a schoolwide discussion about student assessment in the context of curriculum frameworks for specific disciplines, such as math or language arts, then it is highly probable that the dialogue itself might provide powerful staff development that could lead to higher expectations for all students.

ONE SCHOOL'S ATTEMPT TO RAISE STANDARDS

The brief saga of one isolated school's attempt to apply the national frameworks and standards as a way to raise student achievement in the absence of a statewide or national system, might be instructive. Stillwater is a small rural school (600 students, K-6) in a low-income area north of Albany, NY. There, in 1992, a school principal with vision was the catalyst for reform. The principal studied the national assessment frameworks and the drafts of the national curriculum standards. He planned a year-long series of faculty discussions organized around specific grade levels. At first, he encouraged teachers to talk about the K-12 frameworks as an overview for graduation expectations and then debate their potential relevance to their own classes. He urged them to argue about expected outcomes for students, to decide what the second-grade students at their school ought to be able to do and what the third-grade teacher could expect.

At first, the teachers were too reluctant to share ideas. For years the faculty room had buzzed with private accusations that "Mrs. Jones" never taught addition facts in her first grade, or that "Mr. Smith" had totally ignored entire textbook sections on adjectives so that the fifth graders couldn't write vivid descriptions.

After months of grade level meetings, the principal noticed that faculty dynamics began to change. Blame began to be replaced with curiosity and tolerance. Obdurate old-timers began to listen to ideas from younger colleagues. Teachers began to request time to visit one another's classes. One, who had quietly been trying out examples from the National Council of Teachers of Mathematics standards with her own students, gained the confidence to share her results with colleagues. Another solitary fifth-grade teacher, who had pioneered the use of NAEP writing portfolios and scoring rubrics, felt the climate was now "safe" to share her problems and successes. She began to gain respect as a local "expert." Morale improved for many. Several teachers requested sponsorship at summer curriculum training tied to aspects of national standards in the language arts. The school was beginning to build a team approach to a common understanding of what higher standards for students might mean to revising the way they taught.

Then came a sudden halt. The local School Board of Education became anxious. The declining tax base in the community, always shaky, began to erode during a statewide recession. Teacher contract negotiations ran into a brick wall of taxpayer resistance. Work-to-rule set in. The principal decided to move on to an environment more friendly to curriculum reform. The fledgling efforts to bring about change in that school collapsed.

This glimpse of one school's experience three years ago underscores the fragility of the reform process centered on standards. The initiative foundered when their leader left. Another major reason may have been that the New York state mandates and leadership on curriculum and assessment frameworks had not yet crystallized. Change seems unlikely without all these elements necessary for success: robust national frameworks, vision and leadership of the principal, close involvement of teachers, supportive participation from the local community, and reinforcement from statewide commitment to common standards.

DISTRICT REFORM IN A STATEWIDE CONTEXT

Might it be possible to combine the clarity of national standards, the clout of state mandates, and the energy of local practitioners? The past two years' experience of the Cherry Creek School District in Colorado shows a promising pathway. The state context was important. A few years ago, Colorado mandated that each district set standards for grades 4, 8, and 12 in core content areas (Colorado State Department of Education, 1991). The district reform began with a few visionary leaders who saw value in the definition of "basic," "proficient," and "advanced" levels of achievement as delineated by the NAGB reports. They set about to revolutionize the school climate by presenting these definitions as jumping-off places to spur discussion among the entire staff and community. They forged a powerful link between a vigorous site-based reform strategy and the Colorado state mandate to develop standards.

Their elementary executive director, Dr. Barbara Randall (1994), explained that, as a first step, several district people analyzed the NAGB assessment frameworks and the drafts of the national curriculum standards in the core areas of math, science, and language arts. With a few district teachers, curriculum leaders prepared a rough draft of what Cherry Creek students ought to know and be able to do at grades 4, 8, and 12 in those subjects. They then organized a year-long campaign to galvanize teacher ownership and to gain support for exploring techniques and strategies to move students to reach those high standards. They acknowledged that the standards were only a draft and would be revised based on teacher input. Many teacher committees were formed and given extensive time to discuss the standards and their needs to help students achieve them. Teachers expressed doubts. Extended time for discussion was crucial. Their disagreement was respected. There was no expectation that teachers would be forced to change their thinking overnight. At first, the actual definition of essential components of the content was not the focus. The main thrust was the development of a common plan.

The most important part of the process was considered to be the intensive and ongoing teacher conversations that the plan set into motion. After the end of the first year, school leaders reported that the entire climate of the school had been revolutionized. As they enter the third year of the process, they believe that even without final definition of their curriculum frameworks, teachers and students have already begun to reap the benefits in more thoughtful teaching and evidence of student learning.

The Cherry Creek experience suggests how linkage between nationally derived student performance levels, a context of state mandate, and adaptation to local practices might be a useful strategy in the campaign to improve student learning. Critically important elements were visionary school curriculum leaders, respect for thoughtful dialogue, and the kind of commitment of extensive time and money to professional development which Jamentz (1994) recommends.

CURRICULUM STREAMLINING AS A RESULT OF ACHIEVEMENT STANDARDS

As shown in Cherry Creek, the emphasis on achievement levels and high standards for all students could lead teachers and administrators to conduct a realistic reappraisal about what should be taught. In the future, all areas of the curriculum could be affected because high expectations in every subject will require some winnowing. When subjects such as bicycle safety, Arbor Day, drug awareness, and other currently required topics are competing for time with learning about the Civil War, writing essays, or applying math formulas, something has to give. If performance measures of achievement are included in a school assessment system, students will need time in the school day, as well as in homework, to reflect and develop portfolios and "exhibitions." It is probable that some subjects will have to be eliminated (Wiggins, 1987).

Perhaps more schools will follow the lead of Deborah Meier's (1994) example at Central Park East where the staff designed a daily schedule that relegated all noncore subjects to time before or after the block of time reserved for the regular academic core.

In order to build understanding and support in the larger school community for new assessment standards, shared decision-making mechanisms will need expansion (Baker & Linn, 1993). This activity will probably absorb extra staff leadership time and mean reorganization of how the business of school operates. It will require time and commitment to explain proposed changes to parents and other community members. Curriculum directors will have to allocate much more time for planning and supervision as Glickman (1990) warns. Less time will be available for other responsibilities, and changes in instructional schedules will be inevitable.

The move to establish new local and state standards for student performance will undoubtedly lead to greater scrutiny of textbook adoption policies and teaching materials. The development and publicizing of national curriculum standards documents in major subjects, such as math, language arts, history, and the arts, could generate much healthy intellectual discussion among teachers provided the documents find their ways into schools. As teachers deliberate about different ways for students to display competence in a subject, opportunities for new choices about appropriate materials will arise. The NAEP reading assessments in 1992 and 1994 exemplified some nontraditional, challenging ways to tap proficiency in different reading strategies. Those tests paralleled the momentum observed among literature-based reading advocates who advocated the selection of a wide range of "authentic" texts. Similar changes can be detected in other fields. In the area of mathematics, there has been a surge of new books on problem solving. Work done by the National Center for History in the Schools to develop the use of actual artifacts, more primary documents, and simulations to bring issues of the past back to life has begun to affect the social studies curriculum in elementary classrooms.

Some see a potential problem with discipline-based achievement standards, worrying that, if the achievement levels in one subject are developed only in discipline-specific terms, potential interdisciplinary connections might be ignored. They fear that performance assessment narrowly conceived might lead to learning's becoming compartmentalized, trivialized, or less integrated among all the subjects (Baker, 1994; Guskey, 1994). A limited picture of what a student actually knows might emerge. Wiggins (1993b) and others argue persuasively that knowledge of how students fit all the pieces together is needed in order to form an accurate picture of what they know. Champions of opportunity-to-learn standards, such as Darling-Hammond (1991), argue that determining achievement levels in any subject without factoring in the social or academic context in which a student operates may lead to incomplete data and unfair comparisons.

Scholars such as Gordon (1992) fear that if teachers curtail their instruction to focus on that which they can clearly assess, the new assessment might lead to less coverage of important diversity issues and a narrower curriculum.

POTENTIAL CHANGES IN THE GRADING SYSTEM

Widespread use of achievement levels from large-scale assessments could mean more than cosmetic changes in the school grading system. Teachers have always used many different ways to determine whether and to what degree a student has mastered a concept or a skill. The degree to which any new assessment design encourages multiple ways to gather evidence about student skills will be a factor in the acceptance of standards. The introduction of "portfolios, products and performances" into a system of student assessment, while welcome to some teachers, may scare the less experienced who have always relied on "tests" (Wiggins, 1993a; Mitchell, 1992). It will also involve enormous investment of teacher time.

If a system of higher standards for achievement were interpreted to mean that grading policies in classrooms had to be standardized, some analysts, such as Wiggins (1991), predict trouble. Shanker (1994a) recently proposed that, if teachers adopted a policy of assigning standardized grades according to strict state or national criteria, in some schools all the students would get failing grades. He suggested that instead, locally-determined grades could coexist within a system that includes common standards. He argued that, in fact, "all A's are not equal, nor should they be" (p. E7). A grade, he argued, should represent a variety of purposes: to provide incentives to students and to reward effort or improvement (even if the level of work is not world-class). Many advocates of performance assessment (Stiggins, 1988; Wiggins, 1993a) agree that it is desirable for superiority and failure to be relative to a school and a student's past achievement. If that philosophy about the relativity of classroom grades prevails, where will national standards fit in? If all *A*s should not be equal, how then will Sarah's and Cedric's questions be addressed in our schools? How do we ensure that they can be confident that their skills and knowledge will be comparable to world-class standards?

COMMUNITY ENGAGEMENT WITH PERFORMANCE STANDARDS

As noted by a recent Public Agenda Foundation (1994) report, more effective "public engagement" will be needed to ensure success in reforming curriculum and standards. It is particularly difficult for the public to accept and absorb new grading practices, because so much of what people expect of school is based upon their own experiences when they were in school (Urbanski, 1993). The public becomes wary when researchers point to samples of reading tasks within the reach of second graders 40 years ago that are now deemed too difficult for most fourth graders (Chall, 1991). New systems of assessment will probably require students to generate more "constructed responses" to test questions, more student-designed products, and "portfolios" to chronicle the growth of their skills in a variety of subjects. New assessments, however, will not automatically bring about higher achievement. Parents will need help in accepting the validity and necessity of these new ways to gather evidence of student learning. The public will have to be invited into school to see and discuss samples of actual student work.

Some attempts to apply standards may backfire and bring unanticipated results. While many advocates of raising standards are deeply concerned that the least prepared students with fewest resources were discriminated against in the old conventional testing system, they may also be dismayed by results of newer assessments. Shanker's May 1994(b) commentary on the Ohio experience of setting common standards statewide noted that the failure of many minority students to perform adequately has led to great disappointment. The standards imposed at grade 9 in Ohio, even though not particularly rigorous, were in danger of being rescinded because of equity concerns. Similar fears of litigation about fairness of higher standards may create barriers to the widespread use of high standards.

Another obstacle to wholehearted adoption of achievement levels may arise from the problems associated with the growing practice of including children with severe handicapping conditions in the regular classroom. Teachers have found that they rarely receive all the support required to give each child individual help at their instructional level. Focusing on standards is difficult for a teacher who is concentrating on behavior management. It is hard to find the time for standards when teachers are busy translating their lessons into many parallel levels to match all learners' best learning channels.

Without thorough and recurring chances for professional discussions about value-laden issues, teachers may resent standards that they see as unfairly designed to exclude some students from success. They may see standards as obstacles or hurdles for children to struggle over in addition to the daily challenges of no breakfast, no one home after school, and no safe place to play or do homework. To some teachers and parents, standards expecting students to initiate projects, write open-ended essays, and create their own portfolios could be seen as a way to exclude students from their familiar culture. This strong sense of fairness could create resistance from teachers who do not want to risk inflicting more chances for failure on children who are already burdened by poverty.

CONCLUDING THOUGHTS

There are many threats to a thoughtful and fruitful approach to standard setting on a large-scale basis. One of the original uses of the word "standard" was based on medieval knightly battles when the standard was a kind of pennant designating whose team one was on. Policymakers who are not mindful of the divisive consequences of alienating teachers and the public may indeed be doomed to irrelevance. We don't need a battlefield of different pennants; we need to fly one banner. Without careful acknowledgment of the importance of teachers as vital players in the "shared decision-making process," efforts to improve student learning and performance will be limited to rhetoric. Top-down imposition of expectations by policymakers removed from daily classroom obligations could provoke resentment and be rejected as irrelevant to what really goes on.

The real challenge for higher student achievement is how to get behind the closed door of the classroom to help teachers change their beliefs and practices with students. Translating high standards into everyday teaching and learning will require time and practice. Imposing new standards from the outside will threaten many teachers. The fear of high stakes consequences devolving from unfavorable comparisons of their students to others may only cause greater resistance. Unless the teachers have many opportunities to explore adaptations of standards to their own students, they will probably remain stuck in their old habits.

It is my conviction that student performance standards are critical navigational aids to help plot the journey toward improving classroom practices. While the technical tools for determining the North Star of desired expectations for all students must continually be improved, many other factors must also be considered in order to energize the travelers on this journey. Unless teacher beliefs and school culture are changed, strong and effective curriculum leaders in each school sustained, and a system of strong state leadership and well-crafted national assessments aligned with cogent curriculum frameworks put in place, the reform ships may not survive the first leg of the journey. Standards can show educators where to find the North Star, but they cannot power their boats.

The current disjointed relationship between the official, taught, tested, and learned curricula must be addressed. Until all of these forces are understood, little change will be made in what actually happens in most classrooms. National advocacy for reform in standards could be forced into a dormant state until public protest launches the next wave of reform. We may not be able to afford to wait. It would not be fair to Sarah or Cedric.

References

Baker, E. L. (1994). Making performance assessment work. *Educational Leadership, 51* (6), 58-62.

Baker, E. L., & Linn, R. L. (1993-94, Winter). *Towards an understanding of performance standards*. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Bourque, M. L., & Garrison, H. H. (1991). *The levels of mathematics achievement:* Vol. I. National and state summaries. Washington, DC: National Assessment Governing Board.

Cannell, J. J. (1989). *How public educators cheat on standardized achievement tests.* Albuquerque, NM: Friends for Education.

Chall, J. (1991). *Are text-books too easy?* New York: Teachers College Press.

Cheney, L. V. (n.d.). *American memory*. Washington, DC: U.S. Government Printing Office.

Cheney, L. V. (1991). *National tests: What other countries expect their students to know*. Washington, DC: National Endowment for the Humanities.

Cizek, G. J. (1993). On the disappearance of standards. *Education Week, 13*(19), 32, 24.

Colorado State Department of Education (1991). *Colorado sample outcomes and proficiencies for elementary, middle and high school education*. Denver, CO: (ERIC Document Reproduction Service No. PS 020 429).

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58(4),* 438-481.

Cuban, L. (1993, October). The lure of curricular reform and its pitiful history. *Phi Delta Kappan, 75*(2), 182-185.

Darling-Hammond, L. (1991). The implications of testing policy for quality and equality. *Phi Delta Kappan, 73,* 220-225.

Darling-Hammond, L. (1992, November). Reframing the school reform agenda: New paradigms must restore discourse with local educators. *The School Administrator,* 22-27.

Eisner, E. (1994, May). Do American schools need standards? *The School Administrator,* 8-15.

Finn, C., & Ravitch, D. (1994). *Looking back, thinking ahead: American school reform 1993-1995.* Indianapolis, IN: Hudson Institute, Educational Excellence Network.

Flanagan, A. (1994). ELA teachers engaged in massive portfolio project. *The Council Chronicle of the National Council of Teachers of English, 4*(1), 1, 4-5.

Gardner, D. P. (1983). *A nation at risk: The imperative for educational reform. An open letter to the American people. A report to the nation and the Secretary of Education*. Washington, DC: U.S. Government Printing Office. (ERIC Document Reproduction Service No. SP 022 181).

Gentile, C. (1994). *The 1992 writing report card*. National Assessment of Educational Progress. Washington, DC: U.S. Department of Education.

Glickman, C. D. (1990). *Supervision of instruction: A developmental approach*. Boston: Allyn and Bacon.

Gordon, E. W. (1992). *Implications of diversity in human characteristics for authentic assessment*. (Report No. CSE-TR-341). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing. (ERIC Document Reproduction Service No. TM 018 755).

Guskey, T. R. (1994). What you assess may not be what you get. *Educational Leadership, 51*(6), 51-4.

Hewitt, G. (1995). *A Portifolio Primer*, New Hampshire, Heineman (in press), p. 154.

Jamentz, K. (1994). Making sure that assessment improves performance. *Educational Leadership, 51*(6), 55-57.

Lachat, M. (1994). High standards for all: Opportunities and challenges. Final draft of report by the Center for Resource Management for the NH Special Education Program Improvement Partnership. Concord, NH.

Meier, D. (1994, May 8). Changing a tire while the car is moving. A [speech]. Sponsored by the Capitol Area School Development Association, Albany, NY.

McMillan, D. L., Semmal, M. I., Gerber, M. M. (1994). The social context of Dunn: Then and now. *The Journal of Special Education, 27(4)*, 466-480.

Mirel, J., & Angus, D., (1994, Summer). High standards for all? The struggle for equality in the American high school curriculum, 1890-1990. *American Educator*, 4-42.

Mitchell, R. (1992). *Testing for learning: How new approaches to evaluation can improve American schools*. New York, NY: Free Press.

Mullis, I. V. S., Owen, E. H., & Phillips, G.W. (1990). *America's challenge: Accelerating academic achievement (A summary of findings from twenty years of NAEP)*. Princeton, NJ: Educational Testing Service.

New Standards Project. (1994, July). Teachers draw inspiration, perspiration at summer meeting. *The New Standard, 2*(7).

New York State Board of Regents. (1994, June). Curriculum and assessment. [Staff paper]. Albany, NY.

BEST COPY AVAILABLE

Office of Research. (1994). What do students' grades mean? Differences across schools (Res. rep.) Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

O'Neil, J. (1993, February). On the new standards project: A conversation with Lauren Resnick and Warren Simmons. *Educational Leadership, 50*(5), 17-21.

O'Neil, J. (1994). Making assessment meaningful. *ASCD UPDATE, 36*(6), 3-5.

Owens, J. E. (1988, August). *Curriculum and evaluation standards for school mathematics: Report of the National Council of Teachers of Mathematics' Commission on Standards for school mathematics*. Paper presented at the summer workshop of the Association of Teacher Educators, Starkville, MS. (ERIC Document Reproduction Service No. SP 030 756).

Public Agenda Foundation. (1994, September). Effective Public Engagement. [Summary]. *Goals 2000 Community Update, 15*(1).

Randall, B. (1994, September 23). Interview with the author.

Redfield, D. L. (1989, October). *The role of student outcomes in dual purpose teacher evaluation systems: A model for meeting top down and bottom up needs*. Paper presented at the annual meeting of the American Evaluation Association, San Francisco. (ERIC Document Reproduction Service No. TM 017 254).

Resnick, L. (1992, Winter). Standards, assessment, and educational quality. *Stanford Law and Policy Review, 4,* 53-59. (Reprinted in *The Journal for the Education of the Gifted, 17*(4) (Summer, 1994), 401-420).

Romer, R. (1994). How good is good enough? *Goals 2000 Community Update, 15*(1).

Ross, W. (Ed.). (1992). *Teacher personal theorizing: Connecting curricular practice, theory and research*. Binghamton, NY: SUNY Press.

Shanker, A. (1994a, September 11). All *A*'s are not equal. *The New York Times,* p. E7.

Shanker, A. (1994b, May 2) "Standards in Ohio." *The New York Times*, E7.

Simmons, W., & Resnick, L. B. (1993, February). Assessment as the catalyst of school reform. *Educational Leadership, 50*(5), 11-15.

Spillane, R. R. (1993, June 2). Student-achievement standards: Why we need them, why we don't have them, and how to get them. *Education Week,* p. 36.

Standards and assessment: The push for new standards provokes hope and fear--and both are justified. (1993). *Harvard Education Letter, 9*(5), 1-6.

Stiggins, R. (1988). Revitalizing classroom assessment: The highest instructional priority. *Phi Delta Kappan, 69*(5) 363-368.

Suskind, R. (1994, September 22). Poor, black and smart, an inner city teen tries to survive M.I.T. *The Wall Street Journal,* pp.1, 6.

Tyler, R. (1949). *The basic principles of instruction*. Chicago: University of Chicago Press.

Urbanski, A. (1993, September 6). [Speech]. Presented to the Warren, Saratoga Washington Board of Cooperative Services, Saratoga, Springs, NY.

Wiggins, G. P. (1987, Winter). Creating a thought-provoking curriculum: Lessons from whodunits and others. *American Educator, 11*(3), 10-17.

Wiggins, G. P. (1991). Standards, not standardization: Evoking quality student work. *Educational Leadership, 48*(5), 18-25.

Wiggins, G. P. (1993a). *Assessing student performance*. San Francisco: Jossey-Bass.

Wiggins, G. P. (1993b, November). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 75*(2), 200-214.

The Impact of Standards on How Teachers Behave in the Classroom: Promises and Perils

Summary of Break-out Session[1]

The discussion began with the question: *How do all the efforts to set standards at the national and state levels fit together?* The facilitator pointed out that while national standards are a necessity, the participation of states in the process is critical. She cited the efforts of the State of New York in involving their local districts, including the teachers, in the process of determining and agreeing on what students ought to know and do.

Aldrich emphasized the critical role that the states are playing in providing leadership for the local educational agencies (LEAs) in terms of what to implement and in coordinating efforts with other LEAs and other states. The group was able to provide examples from the various states. The discussion extended to the impact of performance standards on local educational policy. Some participants noted the impact of national standards on textbook selection, curriculum integration, and on reorganization in staff development; others shared their anxiety about the return to skills grouping versus ability grouping.

The group noted that states, with their unique history and tradition of interacting with their LEAs, demonstrate differences in management and control of the districts. Many cases were cited: for example, Hawaii, where there is only 1 district; Florida, which has 67 school districts; and Midwestern states, where there may be more than 1,000 districts. Additionally, the group examined variations in implementation of educational reform, including standard setting, for example: (a) *Florida's restructuring* places responsibility on school-based committees for the attainment of school goals; and (b) in *North Dakota*, parents are not represented on committees to construct the test frameworks. The group concluded that whatever the differences in states and their relationship with their LEAs, the primary need in implementing standards is educating staff and the community. Participants cautioned that in many states there is no money to monitor the implementation of any initiatives in the districts.

One participant, referring to Hambleton's presentation, pointed out that translation of classroom practices into performance standards should be a priority. However, other participants, reflecting on their past experiences, indicated that (a) in many circumstances the complexity of learning and teaching at the classroom level is not easily translatable into setting standards and/or meeting standards; and (b) although changing the emphasis from basic to higher order thinking skills may have been accomplished, it is still necessary to work with teachers on content standards. The group agreed that there is a need for a body of teachers to become involved in the movement toward standard setting to complement the work being done by the experts. It was noted that where there is no financial support, effort, as in the case of California, could be defeated. The group noted that at the local level, there is need for action on the part of school boards, which can participate in educating the community about standards.

---

[1] This is a summary of the break-out discussions for Phyllis W. Aldrich's presentation. The session was facilitated by Mary Crovo (National Assessment Governing Board) and recorded by Patricia Dabbs (National Center for Education Statistics).

The discussion again turned to the implementation of standards at the state level. Some participants submitted the view that best practices from one state may not be transferable to another because the needs are so different. The group cited states with populations that include a high number of non-English speaking students, those whose concern is for the lack of talented teachers needed to work in schools for the Native American population, and those that lose their best students because of the poor economic base. The group then shared some practices that have worked for them: workshops for parents, workshops for teachers, looking to the international community to examine different standards, and identifying teaching and instructional strategies that will eventually lead to students' achieving agreed-upon standards.

The presenter was asked about the plans of the Council of Chief State School Officers to provide technical assistance to the states. It was explained that the plan included: (a) explaining content standards, performance standards, and opportunity-to-learn standards; (b) having states define standards; and (c) working with states to evaluate the extent to which these standards are synchronized with national standards. It was noted that federal funding to states is related to development and implementation of their standards.

The Likely Impact of Performance Standards as a Function of Uses: From Rhetoric to Sanctions

Robert L. Linn

Professor of Education and Codirector of the National Center for Research on Evaluation, Standards, and Student Testing, University of Colorado at Boulder

ABSTRACT

Performance standards have many potential uses, ranging from relatively benign reporting mechanisms to the determination of rewards and sanctions. The impact of performance standards will depend upon the specific uses to which they are put. A range of potential uses of performance standards will be identified. Plausible unintended effects, as well as the intended impact of each use, will be discussed using examples from existing assessment programs where possible.

Standards and assessments have been given a central role in the educational reform efforts at both the federal and state levels during the past few years. The centrality of standards and assessments are quite explicit in the Goals 2000: The Educate America Act of 1994 and in both the House and Senate versions of the reauthorization of Title I. Three types of standards--content standards, performance standards, and opportunity-to-learn (OTL) standards--are identified in Goals 2000. The focus of this paper is on the second of these. Consideration of performance standards, almost of necessity, leads to some consideration of assessments that will be used to measure the attainment of levels of achievement defined by the performance standards. Because of the interdependencies among assessments and the three types of standards, occasional observations will also be made about the other two types of standards in Goals 2000, but the focus will remain on performance standards.

More specifically, the comments will emphasize the likely effects, both intended and unintended, of alternative sets of and uses of performance standards. It is argued that the likely impact of performance standards will depend heavily on the uses to which they are put. In particular, variations in the specificity of actions that are based on performance standards and the level of stakes that are attached to the results for educators or for students will greatly influence the nature and magnitude of the effects of implementing performance standards.

DEFINITIONS

Before beginning a consideration of alternative uses of performance standards and the likely impact of those uses, the definition of performance standards, their distinguishing characteristics, and their close connection to content standards deserve comment. As was noted in the report of the National Education Goals Panel (NEGP), Goals 3 and 4 Technical Planning Group on the Review of Educational Standards (1993), there has not always been a clear agreement about the definitions of "content" and "performance" standards. The Goals 2000 Act and the report of the Goals Panel Technical Planning Group have helped provide clarification, but their distinction deserves reiteration.

Content standards are expected to define what should be taught and what should be learned. Although the *Curriculum and Evaluation Standards for School Mathematics* published by the National Council of

Teachers of Mathematics (NCTM, 1988) is still the favorite example of content standards, a plethora of content standards is in various stages of development by other subject matter groups. Performance standards, while dependent on content standards, are distinct. According to Goals 2000, "the term 'performance standard' means concrete examples and explicit definitions of what students have to know and be able to do to demonstrate that such students are proficient in the skills and knowledge framed by content standards" [P.L. 103-227, sec. 3 (a) (9)].

These definitions are consistent with the slightly elaborated ones provided by the Goals Panel Technical Planning Group, which offered the following definition of performance standards:

> *Performance standards*. Performance standards specify "how good is good enough." In shorthand, they indicate how adept or competent a student demonstration must be to indicate attainment of the content standards. They involve judgments of what distinguishes an adequate from an outstanding level of performance. . . . Performance standards are *not* the skills and modes of reasoning referred to in the content standards. Rather, they indicate both the nature of the evidence (such as an essay, mathematical proof, scientific experiment, project, exam, or combination of these) required to demonstrate that content standards have been met *and* the quality of student performance that will be deemed acceptable (what merits a passing or an "A" grade). (NEGP, p. 22)

This definition shows the distinction between performance and content standards as well as the close link required between the two. The addition of the requirement that evidence is needed to show that students have met the standards also makes clear the interdependency of performance standards and assessments.

MAJOR USES OF PERFORMANCE STANDARDS

Performance standards have a variety of potential uses, and there is no claim that the four that are distinguished here are either mutually exclusive or exhaustive. Other categorization systems are certainly possible, but the four uses identified are sufficient to illustrate some of the critical differences in both types of use and in likely effects. Four potentially important uses of performance standards are: exhortation, exemplification of goals, accountability for educators, and certification of student achievement. The likely effects of these four major potential uses are quite different.

Exhortation

A major use of performance standards is to exhort educators, students, and the public to exert greater, and possibly different, kinds of effort to achieve established standards of performance. Lesgold (1994) captured the sense of the exhortation role when he likened standards to "the rabbit at a greyhound track--if only we put standards out there people will chase after them" (p. 9). This first function is consistent with the use of "standard" in the sense of a banner or symbol for people to follow.

The exhortation role is an old one. It is also the only one of the four potential uses that requires neither great specificity regarding the levels of performance associated with the standards or specific assessments that can be used to establish unambiguously if they have been attained. That is, performance standards intended only to exhort might not even satisfy the Goals 2000 requirement for "concrete examples and explicit definitions of what students have to know and be able to do."

Although exhortation will undoubtedly continue to be an important use of performance standards, it clearly is not sufficient to accomplish the major improvements in education that are sought. If exhortation were sufficient, it would not be necessary to discuss standards-based reforms. For the past decade, there has been a continual barrage of exhortations based on the argument that performance is not up to standards. Recall the strong language in *A Nation at Risk* (National Commission on Excellence in Education, 1983) in which the achievement of the nation's students was judged to fall far short of the standards of performance required by an increasingly competitive global economy. Of course, the exhortation role of performance standards did not begin in 1983. As Cremin (1989) noted:

> The argument over standards is surely as old as the world itself. Just about the time Adam first whispered to Eve that they were living through an age of transition, the Serpent doubtless issued the first complaint that academic standards were beginning to decline. The charge of decline, of course, can embrace many different meanings and serve as a surrogate for a wide variety of discontents, only one of which may be that young people are actually learning less. As often as not, it suggests that young people are learning less of what a particular commentator or group of commentators believe they ought to be learning, and the "ought" derives ultimately from a conception of education and of the educated person. (p. 7)

It is not meant to imply here that the exhortation function of standards is unimportant. On the contrary, the symbolic use of performance standards can play a vital role in rallying people to action. The rash of education legislation in the wake of *A Nation at Risk* is a case in point. The effects, however, are at best indirect and may be quite unpredictable because they depend on the nature of actions that are taken in their name.

Exemplification of Goals

A second potentially important use of performance standards is to provide clear specifications of the achievement levels that students are expected to attain. This use of performance standards does not require the attachment of high stakes to the results for either students or teachers. The National Assessment Governing Board's (NAGB) Achievement Levels for the National Assessment of Educational Progress (NAEP) provide an example of such a use of performance standards. Among other things, the levels are intended to specify what students should be able to do and to indicate to policymakers and the public how students perform in relation to those standards.

The utility of performance standards to exemplify goals and to compare actual student performance to those goals depends heavily on the clarity and accuracy of the communication of the standards. Much of the controversy surrounding the NAGB Achievement Levels (see, for example, American College Testing, 1993; Burstein et al., 1993; Shepard, Glaser, Linn, & Bohrnstedt, 1993; Stufflebeam, Jaeger, & Scriven, 1991) stems from concerns about the clarity and accuracy of the communication about actual student achievement that the achievement levels show. Although both the step of converting a description of a performance standard into a defined range of scores on an assessment and the step of describing what students who meet a given standard of performance actually know and can do may appear straightforward, satisfactory accomplishment of these two steps has proven to be extraordinarily difficult. Both steps are essential if student achievement is to be reported clearly and accurately in terms of performance standards.

The potential impact of the exemplification role of performance standards is profound and will require considerable time and effort for the full potential to be realized. Clear examples of actual student performance that is judged to meet the high levels of accomplishment of the knowledge and skills defined by the content standards can be powerful material for use in staff development. Although anecdotal, consistent reports from teachers brought together to score responses to extended-response performance tasks support this claim.

For performance standards to have the desired impact on day-to-day classroom activities, they must be internalized by teachers and students. At a minimum, such internalization requires multiple examples of actual student work that is judged to meet particular standards of performance. Because the vision of both content standards and performance standards that are being advanced is a radical departure from the vast majority of current daily classroom practices, the transformation cannot be expected to take place overnight. The progress that has been achieved as the result of concerted efforts of proponents of the *Curriculum and Evaluation Standards* of the NCTM (1988) during the past five or six years demonstrates that the task is both difficult and time consuming.

Accountability for Educators

Systems of accountability for educators relying on the use of assessments and performance standards have been introduced in a number of states. As the Kentucky Educational Reform and Accountability Act of 1990 illustrates (see, for example, Guskey, 1994, p. 18), high stakes may be attached to the results for educators but not to those for students. In Kentucky, schools, not students, can obtain rewards or sanctions depending on the progress achieved in meeting performance standards.

Earlier test-based reform efforts provide considerable experience. Admittedly, there are some important changes in the current efforts to use assessments and performance standards as tools of reform. The explicit use of performance standards distinguishes current efforts from previous accountability systems that focused on scores on norm-referenced achievement tests. The nature of the assessments has also changed. Performance assessments are emphasized, albeit not exclusively, rather than the familiar multiple-choice tests. Although performance assessments come in many varieties, ranging from short-answer items that look to some students like multiple-choice items without options to tasks requiring extended responses or even to portfolios of student work accumulated over the school year, they all require students to perform or construct responses rather than to select the correct option. Nonetheless, some lessons from earlier test-based accountability efforts are likely to be relevant to current efforts.

WYTIWYG ("what you test is what you get") has become a familiar acronym in the performance assessment movement. This notion that tests shape instruction when stakes are attached to the results for teachers or schools is widely accepted and supported by a reasonable amount of hard evidence (e.g., Madaus, West, Harmon, Lomax, & Viator, 1992; Shepard, 1990; Smith & Rottenberg, 1991). It is a cause for concern to those who worry about tests covering only a narrow range of the intended curriculum and overemphasizing factual knowledge and low-level procedural skills. Proponents of performance standards and performance-based assessments not only accept the notions of WYTIWYG and its converse, "you don't get what you don't test," but they also embrace the concepts. The three principles presented as possible "guidelines from accountability assessments" by Resnick and Resnick (1992), for example are:

1. You get what you assess.

2. You do not get what you do not assess.

3. Build assessments toward which you want educators to teach. (p. 59)

The third principle is a logical conclusion if the first two principles are accepted and if some form of assessment for accountability purposes is almost inevitable, as I believe. It should be noted, however, that it is not clear if the third principle can be achieved without unintended negative side effects. Some of the distortions of instruction and corruption of the indicators that resulted from the high-stake uses of standardized-achievement tests (e.g., Linn, Graue, & Sanders, 1990; Shepard, 1990) are not unique to that form of testing.

A concentration of effort on knowledge and skills judged important and therefore included in the assessments can have a desirable effect on the accomplishment of those goals. On the other hand, no assessment can be expected to adequately represent the full domain of content standards in a given discipline. As Madaus (1988) has documented, corruption of the indicator has been a long-standing problem in other countries where the form of assessment is more in keeping with the current performance-based assessment movement in this country. His example of formulaic compositions written for annual essay examinations in Ireland is only one of many examples in this regard.

Perhaps an even greater negative side effect that can be anticipated concerns the content areas that are not assessed. This concern was recognized by a proponent of performance standards and scales of measurement to determine attainment of those standards from an earlier era. Bobbitt (1913) made the following observation in his plea for standards and scientific measures as educational management tools some eight decades ago:

> It will work harm to establish definite standards for only a portion of the tasks of education, leaving the rest to the traditional vagueness and uncertainty of aim. The remedy is to establish equally definite standards in every field of educational labor. There can be no other remedy. (pp. 42-44, quoted in Callahan, 1962, p. 85)

The "remedy" proved much more elusive than Bobbitt ever imagined. Not only is it more difficult to measure some valued aspects of education, but the array of field[s] of educational labor" is immense. Recognition that not being included in the system of assessments and performance standards of an accountability system relegates a field to second-class status, or worse, has led to the proliferation of efforts to set content standards for which it is hoped assessments and performance standards will be added later. How many discipline-based content and performance standards can schools reasonably be held accountable for? The struggle over what gets emphasized, what gets included, and what gets excluded from the assessments and performance standards that count is a struggle over educational values. As Cremin (1989) noted, "standards involve much more than determinations of what knowledge is of most worth; they also involve social and cultural differences, and they frequently serve as symbols and surrogates for those differences" (p. 9). The attempt by the Traditional Values Coalition to have the Alice Walker story "Roselily" removed form the California Learning and Assessment System because

BEST COPY AVAILABLE

the Coalition claimed the passage undermined religious values (Asimow, 1994) is just one of many illustrations of the type of struggle over educational values referred to by Cremin.

In addition to struggles between disciplines and the priorities and realizations within disciplines, accountability by means of assessments and performance standards faces another challenge in the form of proponents of a more integrated curriculum. The emphasis on integrated skills that cut across traditional disciplinary lines found in the SCANS report from the U. S. Department of Labor (1991) and in the Blueprint 2000 educational improvement effort in Florida (Florida Commission on Education Reform and Accountability, 1993) present a view that is quite different from either the current credit structure in high schools or the disciplinary imperialism that is likely to be fostered by some uses of the emerging content standards.

Certification of Student Achievement

Clearly the potential use of performance standards with the highest stakes is student certification. This might be in the form of graduation requirements, endorsed diplomas, or special certificates. If performance standards are used for determining employment opportunities, eligibility for college, or other desired outcomes for individual students, however, both the strength of the potential impact and the technical requirements for assessments and performance standards will be raised.

There are existing examples of performance standards and assessments that are used for student certification. The New York Regents Examinations, for example, have a long history that is relevant here. Other examples include the California Golden State Examinations and the College Board Advanced Placement (AP) courses and examinations. The widespread use of examinations for student certification in other countries is also relevant. The many minimum competency testing programs that were introduced around the country in the 1970s and early 1980s provide yet another example.

Although much can be learned from these examples of student certification uses of performance standards and assessments, there are some features of the current movement that are unique. Unlike minimum competency requirements, for example, the current press for performance standards emphasizes high standards rather than minimum standards. In contrast to the other examples such as AP, the current emphasis is on all students rather than on only college-bound and higher achieving students. These are important distinctions that together with other demands create some serious tensions to be discussed later.

Madaus (1988) effectively made the case that when high stakes are attached to the results of an examination, "a tradition of past exams develops, which eventually de facto defines the curriculum" (p. 93). He illustrated this with an example related to him by Ben Bloom based on his observation of a former student teaching in his classroom in India. When the former student began to discuss a broader set of implications related to the topic of the lesson, the students began to chant "NOE, NOE," an acronym familiar to them and their instructor for "not on the exam." This illustrates that in the extreme there are likely to be dysfunctional aspects of assessments when the stakes for individual students become possible even if it proves possible to "build assessments toward which you want educators to teach" (Resnick & Resnick, 1992, p. 59) because the intended curriculum is always broader than any assessment.

Concern that high stakes would be attached to assessment results for individual students provided at least part of the motivation for the third category of standards in Goals 2000. Opportunity-to-learn standards are expected to define criteria for assessing the adequacy of the learning opportunities provided for students. Although both content and performance standards can be the source of some controversy, they are far less controversial and contentious than OTL standards. The opposing views of OTL standards were recently summarized by Porter (1994) as follows:

> To proponents, OTL standards represent the age-old problems of equity in education. In particular, advocates of OTL standards see them as an appropriate antidote to the potentially negative effects of high stakes testing on students who, through no fault of their own, attend schools which provide an inferior education. To opponents, OTL standards evoke all their worst fears about federal intrusion into local control of the quality and nature of education. (p. 1)

Even without the inclusion of OTL standards in Goals 2000, the issue of opportunity to learn would surely be raised in any high-stakes use of performance standards for student certification. As Philips (1994) will undoubtedly discuss in this panel presentation, with the *Debra P. v. Turlington* (1981) case as a precedent, OTL is apt to be an issue in any legal challenge to the student certification use of standards.

CONCLUDING COMMENTS

Together with content standards, performance standards are expected to contribute to the accomplishment of a great deal. The standards are expected to meet multiple demands. These demands are sometimes conflicting and therefore create tensions. Five expectations for performance standards are particularly notable in this regard. The standards are expected to:

1. Be set at high levels that are "internationally competitive and comparable to the best in the world" [P.L. 103-227, sect. 213 (a) (2) (A) (I)];
2. Have a small number of levels (e.g., advanced, proficient, not yet proficient);
3. Apply to all students;
4. Be used in all content areas; and
5. Allow for integration across disciplinary lines.

The minimum-competency standards set in a number of states applied to almost (but not literally) all students, but they were neither high nor were they invoked in many content areas. AP standards are relatively high, but they do not apply to even a majority of, much less all, students. Furthermore, they are tied to specific courses and course syllabi, not to universal requirements for graduation or to integration of knowledge and skills across disciplines. Even so, they allow for five levels, not three.

The emphasis on the education of *all* students is not just politically correct, it is right to do for both moral and pragmatic reasons. The rejection of normative notions that relegate half the students to below-average performance in favor of assessing performance in terms of fixed standards also has considerable power and appeal. Moreover, there are plenty of good reasons for supporting the high-standards side of the reform efforts. There is a tremendous gap between the rhetoric of high standards for all students and the current state of affairs, however. The concern for opportunity to learn in schools barely scratches the surface of what is needed.

Immediate implementation of high performance standards with distinctions made only for a small number of levels would yield politically unacceptable levels of failure, resulting in pressures that are likely to have one or more undesirable effects such as lowering the standards, corrupting the indicators, distorting instruction, or abandoning the effort in favor of a new educational fad. High-stakes accountability for educators rather than students will create somewhat less pressure on the system, but careful attention should be given to monitoring both unintended and intended effects to ensure that the positives outweigh the negatives.

In my judgment, the most positive impact of performance standards is apt to come from their use in exemplification of goals. This is admittedly not the quick fix that is likely to appeal to many politicians, but in the long run, real educational reform requires the kind of professional development that can be accomplished only with time and considerable effort.

References

American College Testing. (1993). *Setting achievement levels on the 1992 National Assessment of Educational Progress in mathematics, reading, and writing: A technical report on reliability and validity*. Iowa City, IA: Author.

Asimow, N. (1994, February 25). Alice Walker story furor grows. *San Francisco Chronicle*, p. A20.

Bobbitt, F. (1913). *The supervision of city schools. The twelfth yearbook of the National Society for the Study of Education*. Chicago: University of Chicago Press.

Burstein, L., Koretz, D. M., Linn, R. L., Sugrue, B., Novak, J., Lewis, E., & Baker, E. L. (1993). *The validity of interpretations of the 1992 NAEP achievement levels in mathematics*. Los Angeles: UCLA Center for the Study of Evaluation.

Callahan, R. E. (1962). *Education and the cult of efficiency: A study of the social forces that have shaped the administration of the public schools*. Chicago: University of Chicago Press.

Cremin, L. A. (1989). *Popular education and its discontents*. New York: Harper & Row.

Debra P. v. Turlington, 474 F. Supp. 244 (M.D. Fla. 1979), *aff'd in part, rev'd in part*, 644 F2nd 397 (5th Cir. 1981); *on remand*, 564 F. Supp. 177 (M.D. Fla. 1983), *aff'd*, 730 F2d 1405 (11th Cir. 1984).

Florida Commission on Education Reform and Accountability. (1993). *Blueprint 2000: A system of school improvement and accountability. For 1994-95 school improvement plans*. Tallahassee, FL: Florida Department of Education.

Goals 2000: The Educate America Act, Pub. L. No. 103-227, (1994).

Guskey, T. R., (Ed.). (1994). *High stakes performance assessment: Perspectives on Kentucky's education reform*. Thousand Oaks, CA: Corwin Press.

Lesgold, A. (1994, March). *Standards and assessment: Toward a research agenda for policy and research*. In summary of a workshop convened by the Board on Testing and Assessment. Board Bulletin. Washington, DC: National Research Council.

Linn, R. L., Graue, M. W., & Sanders, N. N. (1990). Comparing state and district test results to national norms: The validity of claims that "everyone is above average." *Educational Measurement: Issues and Practice, 9* (3), 5-14.

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum. Eighty-seventh yearbook of the National Society for the Study of Education, Part I* (pp. 83-121). Chicago: University of Chicago Press.

Madaus, G. F., West, M. M., Harmon, M. C., Lomax, R. G., & Viator, K. A. (1992). *The influence of testing on teaching math and science in grades 4-12: Executive summary* (Tech. Rep.). Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

National Academy of Education. (1993). *Setting performance standards for student achievement.* A report of the National Academy of Education Panel on the evaluation of the NAEP trial state assessment: An evaluation of the 1992 achievement levels. Stanford, CA: Stanford University, National Academy of Education.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: U.S. Government Printing Office.

National Council of Teachers of Mathematics. (1988). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Education Goals Panel, Goals 3 and 4 Technical Planning Group on the Review of Education Standards (1993). *Promises to keep: Creating high standards for American students.* Washington, DC: Author.

Phillips, S. E. (1995). Legal defensibility of standards: Issues & policy perspectives. In *Proceedings of Joint Conference on Standard Setting for Large-Scale Assessments* (pp. 379). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

Porter, A. (1994, May). *The uses and misuses of opportunity to learn standards.* Paper presented at the Brookings Institution Conference on Beyond Goals 2000: The Future of National Standards in American Education. Washington, DC.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. G. Gifford & M. C. O'Conner (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer Academic.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching to the test? *Educational Measurement: Issues and Practice, 9* (5), 15-22.

Smith, M. L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice, 10,* 7-11.

Stufflebeam, D. L., Jaeger, R. M., & Scriven, M. (1991). *Summative evaluation of the National Assessment Governing Board's inaugural 1990-91 effort to set achievement levels on the National Assessment of Educational Progress.* Washington, DC: National Assessment Governing Board, August 1991.

U. S. Department of Labor (1991). *What work requires of schools: A SCANS report for America 2000.* Washington, DC: Author.

The Likely Impact of Performance Standards as a Function of Uses: From Rhetoric to Sanctions

Summary of Break-out Session[1]

This discussion focused on four areas: (a) opportunity-to-learn (OTL) standards, (b) the compensatory nature of performance levels, (c) levels of specificity in the description of student proficiencies, and (d) the impact of data on the standard-setting process. Specific questions guided the discussion in each area.

The first question posed was: *How is OTL operationally defined and how is it related to the National Assessment of Educational Progress (NAEP)*? It was submitted that not all students have an opportunity to learn what is recommended in content standards, for example the National Council of Teachers of Mathematics (NCTM) standards; however, there is an urgency to include those areas in the assessment. There were differing opinions in the discussion of operationally defining OTL standards as well as in discussing whether students should be tested only on the content that they have had an opportunity to learn. It was suggested that any operational definition of OTL would be too specific and counterproductive. The group agreed that all students should have an opportunity to learn all that has been specified. However, it was noted that in content areas like mathematics and science, unlike in history and English, the standards and goals are clear, and, therefore, the assessment is more easily defined. It was also suggested that even if students are assessed on materials that they did not have an opportunity to learn, the results should be reported in two categories: content that students have had the opportunity to learn and those which they have not.

The group then discussed the following questions: (a) *How do the OTL standards relate to what students should know and be able to do*? and (b) *What's the point of gathering background information, such as calculator use*? It was suggested that the OTL standards have less relevance for large-scale assessments that have no consequences for individual reporting than for high-stakes, individual-level assessments. With reference to the second question, it was suggested that background information is needed to evaluate OTL and that the information is not used for accountability.

The discussion then focused on the compensatory nature of performance levels. The question that initiated the discussion was: *Are all students required to achieve the highest level of proficiency across different subject areas or categories of performance tested*? It was noted that since the Goals 2000: Educate America Act calls for high standards for *all* students, there is a need for the development of compensatory models to assist students at the high school level. One participant suggested that in NAEP, the aim is not to categorize students as "advanced," "proficient," etc., based on narrowly defined basic skills but on complex proficiencies; it was emphasized that students are expected to achieve at a high level. The group was of the opinion that there is a need to continue the dialogue related to the setting of achievement levels in NAEP.

---

[1] This is a summary of the break-out discussions for Robert L. Linn's presentation. The session was facilitated by Susan Loomis (American College Testing) and recorded by Sheida White (National Center for Education Statistics).

The discussion then focused on *levels of specificity in the description of what students can or cannot do*. The group agreed that there is a need for descriptions of performance standards and of technical information on testing, written in language that can be understood by the public. It was suggested that parents must receive both information related to assessment outcomes and samples of students' performance. Additionally, it was indicated that parents and the public must be assisted to understand the shift from traditional testing, which focused on isolated specific skills, to performance assessment, which is based on proficiency and is integrated across content areas. It was further suggested that parents' concern about their children's educational progress--moving to the next grade, going to college, getting a job--should also inform the setting of the standards.

The group then discussed the role of impact data on the standard-setting process. The group emphasized the importance of the training of the judges and of establishing balance within the committee; it also acknowledged that some outstanding teachers have made significant contributions to standard-setting panels. It was also suggested that if proficiency levels are set the same for all students, including those in compensatory, special education, and bilingual programs, then the same students will do poorly every time. A change in the design of the performance standards was recommended.

Legal Defensibility of Standards: Issues & Policy Perspectives

S. E. Phillips

Professor, Michigan State University

ABSTRACT

Legal challenges to standards may be expected when standards create adverse impact on historically disadvantaged groups; use processes perceived to be unfair, arbitrary, or capricious; imply that specific attitudes or values will be assessed; fail to consider accommodations for the disabled; or appear to assess knowledge/skills that examinees have not had the opportunity to learn. This paper will discuss the legal framework surrounding these issues, provide guidance on developing policies and procedures that are legally defensible, and consider the types of empirical data that may be relevant in the event of litigation.

BACKGROUND

To obtain planning grants under the Goals 2000: The Educate America Act (1994) (federal legislation for reforming American schools), states or school districts must agree to set high content and performance standards for all students. Large-scale assessments are expected to be implemented to measure attainment of these required standards. Such assessments will have standards of their own in the form of cut scores, which, at a minimum, will classify students as having met or not met the standards. Many states already have such standards at the high school level in the form of required graduation assessments.

In a recent *Education Week* article (Pitsch, 1994), critics of Goals 2000 contend that the standards-based reforms it encourages will erode local control over schools and instill in students state-sanctioned attitudes, values, and beliefs. In response, proponents of Goals 2000 argue that the critics have misinterpreted the legislation. As states and local districts debate whether to seek federal funding, they must be cognizant of the opposition and the potential for political and legal challenges to mandated standards and the assessments used to measure their achievement. Precedents from challenges to state graduation assessments provide some insight into the potential legal arguments that may be raised by parents and other challengers.

HIGH-STAKES ASSESSMENT

Large-scale assessments may be high stakes or low stakes. High-stakes assessments are used to make important decisions about individuals or to distribute monetary rewards. Low-stakes assessments often provide only group data and are used for evaluation and planning.

In the educational context, high-stakes assessments typically report individual student scores and are used to award diplomas, licenses, merit pay or other monetary recognition. Some educators believe that large-scale assessments reported by the media also carry some indicia of high stakes. It is in this context that legal defensibility is most salient.

Administrators of high-stakes, large-scale assessment programs would like a direct answer to the question, "What is required to make the program legally defensible?" Because there are a multitude of ways in which such programs may be challenged, there is no single answer to this question. Answering this question is further complicated for educational performance assessments because as yet there are no specific cases in which courts have enunciated required legal standards. However, there are precedent cases for large-scale, multiple-choice tests in education and for performance assessments in employment and higher education contexts. Thus, it is possible to provide some guidelines for what may be legally required.

TWO TYPES OF STANDARDS

There are two interpretations of the term "standards" that may play a role in the legal defensibility of a high-stakes, large-scale assessment program. These interpretations include: (a) goal statements describing what students should know and be able to do in specific content subjects, and (b) the specification of the score or level of performance corresponding to a category of achievement such as "passing" or "proficient." Both interpretations of the term "standards" are included in the discussions of legal defensibility that follow.

Challenges may be targeted at specific standards statements or at the performance levels required for a particular category of achievement. A standard may be acceptable on its face, but the specific way in which it is assessed or implemented may render the overall process unfair to some individuals and groups. When challengers perceive that direct challenges to objectionable standards may be infeasible or unsuccessful, other aspects of the assessment program may be targeted. Thus, the assessment itself may become a lightning rod for criticisms that originate in disagreement with federal, state, or local school standards and policies.

There are several specific conditions that appear most likely to trigger a legal challenge to a high-stakes, large-scale assessment and its associated standards. They include adverse impact on historically disadvantaged groups; use of processes perceived to be unfair, arbitrary, or capricious; suggestion that specific attitudes or values are being assessed; failure to provide all accommodations requested by the disabled; and assessing knowledge or skills that examinees have not had the opportunity to learn.

Although these issues apply to all types of assessments, there are special concerns to be addressed as programs move from multiple-choice tests to performance assessments. This occurs for a variety of reasons including the lack of experience with performance assessments in subjects other than writing, the lag in the development of defensible technologies and methodologies for scaling and equating performance assessments, and the increase in errors due to lack of generalizability of task samples and the fallibility of human judgments.

PROFESSIONAL STANDARDS

In reviewing prior cases involving challenges to assessments, it is evident that courts apply relevant professional standards as articulated by expert witnesses. Thus, the *American Educational Research Association (AERA)/American Psychological Association (APA)/National Council on Measurement in Education (NCME) Standards for Educational and Psychological Testing* (APA 1985), will likely be cited in any legal challenge to an educational assessment program. However, the *Test Standards* (APA, 1985) were developed when multiple-choice testing was dominant and may not be detailed enough to represent professional consensus for the emerging technologies associated with performance assessments. The revision of the *Test Standards* (APA, 1985) currently in progress may more fully address such issues. The *Code of Fair Testing Practices* (Joint Committee on Testing Practices, 1988) also provides guidance for assessment developers and users and may be cited by measurement experts.

THE *DEBRA P.* REQUIREMENTS

To provide a framework for evaluating the legal defensibility of assessment program standards, it is instructive to examine past cases to determine the ways in which courts tend to approach assessment-related challenges. The landmark case for large-scale educational assessment is *Debra P. v. Turlington* (1979/1981/1983/1984). The *Debra P.* case involved the denial of high school diplomas to historically disadvantaged students who had failed a minimum-competency examination in Florida.

In 1976, the Florida legislature established the Functional Literacy Examination (FLE) as the state's graduation test, effective for the 1979 graduating class. The FLE was a multiple-choice test of basic communication and mathematics skills applied to real-life situations. Graduating seniors who had not passed the test after multiple retakes were awarded a certificate of completion. After three administrations of the FLE, approximately 2% of the Caucasian seniors had not passed while approximately 20% of the African American seniors had not passed.

In addressing the claim that imposition of the FLE discriminated against African American students, the appeals court crafted new legal standards for graduation tests. Two major requirements for diploma testing emerged from the *Debra P.* case: notice and curricular validity.

DUE PROCESS

The notice and curricular validity requirements both emanate from the due process clause of the Fourteenth Amendment to the U.S. Constitution. It states that "[n]o state shall . . . deprive any person of life, liberty, or property, without due process of law. The Court held in the *Debra P.* case that a high school diploma is a property right subject to Fourteenth Amendment due process protection.

There are two types of due process that have been recognized by federal courts: procedural and substantive (Nowak, Rotunda, & Young, 1986). Procedural due process focuses on assessment administration and sets an expectation that the procedures and processes implemented by an assessment program will be fair and equitable. Substantive due process focuses on the assessment instrument itself and sets an expectation that it will follow professional standards, be valid and reliable, and be fair to all examinees. Substantive due process can be violated when the knowledge and skills being assessed are judged to be arbitrary, capricious, or unfair.

NOTICE

Notice is a procedural due process requirement. In the *Debra P.* case, the court held that students who might be deprived of their property rights in a diploma must receive adequate prior notification of any required assessment.

The purpose of the *Debra P.* notice requirement is to ensure fairness by requiring that students be notified well in advance of the implementation of a graduation assessment. Although the court in the *Debra P.* case found a four-year notice period adequate, this time frame also corresponded to elimination of dual schools in Florida. Thus, courts may find a shorter notice period acceptable under other circumstances. However, courts have uniformly declared notice periods of less than two years inadequate (*Anderson v. Banks*, 1982; *Board of Education of Northport-E. Northport v. Ambach,* 1983).

The length of the required notice period probably depends, in part, on the success of assessment administrators in disseminating relevant information about content, format, and scoring to school personnel and students. Although it is probably not necessary to communicate specific passing standards ahead of time, students and school personnel should be provided with clear indications of the specific knowledge and skills for which students will be held accountable and general guidelines on what constitutes acceptable performance. In addition to providing curricular frameworks or assessment specifications, the provision of sample items along with model responses may also be helpful in communicating expectations.

CURRICULAR VALIDITY

The second major requirement that emerged from the *Debra P.* case was curricular validity. The court derived the curricular validity requirement from substantive due process protection.

Curricular validity requires test administrators to demonstrate that students have had an opportunity to learn the knowledge and skills included on a graduation test. The *Debra P.* court accepted survey responses from Florida teachers who rated the tested skills as important and said they taught them as sufficient evidence to satisfy this requirement. The court did not require Florida to demonstrate that each student individually had been exposed to instruction on each tested skill. Overall, the Florida data indicated that on average Florida students had more than two opportunities to learn each tested skill. This result occurred because most tested skills were taught at more than one grade.

The *Debra P.* court held that the relevant evidence for establishing curricular validity was that "the skills be included in the official curriculum and that the majority of the teachers recognize them as being something they should teach" (564 F.Supp. at 186). The collection of a variety of evidence, including teacher surveys, student surveys, and reviews of textbooks and curricular guides, is desirable in establishing curricular validity. Further discussion of the *Debra P.* case, the distinction between instructional and curricular validity, and the issues involved in the closeness of the match between assessments and curricula can be found in Phillips (1991, 1993a).

ASSESSMENT REVISIONS

A related curricular validity issue involves revision of an existing assessment program. Sometimes the groups that develop the curricular frameworks on which graduation assessments are based want to include skills and performance standards that teachers are not currently teaching but which the curricular group believes teachers ought to start teaching. Immediate inclusion of new skills and standards in the assessment may have the desired effect of changing teaching practices but is likely to be viewed as unfair by the courts. Any time new skills or performance standards are added to a graduation assessment, the notice period should probably be as long as that for the implementation of the original assessment. Such a notice period will ensure ample time for incorporation of the new material into the curriculum so that students will have a reasonable opportunity to learn it.

CURRICULAR VALIDITY OF PERFORMANCE ASSESSMENTS

Performance assessments may have additional opportunity-to-learn issues related to format. For example, if writing is a part of science or mathematics assessments but most students have not been required to write in science or mathematics classes, this may indicate lack of curricular validity even if the specific content necessary to answer the questions has been taught. In addition, in cases where complex multiple skills are tested in a single task, it may be difficult to identify the point in the curriculum where students are expected to be able to successfully complete such a task. To the extent that the skills required for success on performance tasks can not be specifically taught in a short time frame, it may also be difficult to demonstrate adequate remediation efforts for students who fail. Further discussion of this and other legal issues related to performance assessment can be found in Phillips (1993a; 1993b).

PRIOR PREDICTION OF PERFORMANCE ON GRADUATION TESTS

One must also be careful about using assessment standards in prior grades to predict success on a graduation assessment. Earlier assessments may test different skills at different levels of competence, and such predictions assume constant student effort and motivation over time. Predicting lack of success and permanently tracking students into remedial programs may be viewed negatively by some educators, while predicting future success may create an erroneous expectation of entitlement to a diploma. Given errors of measurement and prediction, the increasing level of content difficulty, and the lack of control over other factors that might significantly impact success, such predictions are dubious. However, identifying grade-appropriate standards that have not been met and providing remediation early in the instructional process can be beneficial.

CHANCES IN ADVERSE IMPACT

The equal protection clause of the Fourteenth Amendment to the U.S. Constitution provides that "[n]o state shall . . . deny to any person within its jurisdiction the equal protection of the laws." To demonstrate an equal protection violation, federal courts require evidence that one of two similarly situated groups was treated differentially by a state or local government that intended to discriminate (Nowak et al., 1986).

Adverse impact occurs when a government policy results in more frequent denial of a property right to one group than to another. That is, there is adverse impact when substantially more historically disadvantaged students fail a graduation test than nondisadvantaged students. Adverse impact is a necessary but not sufficient condition for establishing intent to discriminate. To evaluate intent to discriminate, federal courts must also consider all the facts and circumstances surrounding the challenged state or local government action (*Village of Arlington Heights v. Metropolitan Housing Dev. Corp.*, 1977; *Personnel Administrator v. Feeney*, 1979).[1]

---

[1] Note that employment assessments, unlike graduation tests, are subject to Title VII of the Civil Rights Act and the corresponding Equal Educational Opportunity Commission (EEOC) *Uniform Guidelines for Employee Selection*, which provide for a presumption of discrimination when the success rate for disadvantaged applicants is less than 80% of that for nondisadvantaged applicants.

The *Debra P.* case focused specifically on the initial adverse impact of the Florida test on African American students. However, after four years of litigation, with remediation and retesting the gap between the performance of African American and Caucasian students had narrowed from a difference of approximately 18% to a difference of less than 10%. In general, in the last decade, the gap between nondisadvantaged and historically disadvantaged performance on multiple-choice tests has seemed to narrow. After remediation and multiple retakes, historically disadvantaged students' passing rates on graduation tests have been relatively high and much closer to nondisadvantaged students' passing rates than when such tests were first adopted (*Debra P.*, 1979-1983, Phillips, 1993b).

With new standards and performance assessments, differential passing rates will again receive careful scrutiny. Although differential performance by itself is not sufficient to invalidate an assessment program, defending such a program against a legal challenge based on alleged discrimination can be costly, time-consuming, and detrimental to public relations. It also can exacerbate test security concerns as challengers seek access to the disputed assessment tasks.

Because performance tasks tend to measure complex sets of skills that combine reading and writing with subject specific knowledge, students who are disadvantaged may perform less well. Initial data suggest that there is increased adverse impact of new standards and performance assessments on historically disadvantaged students (Beck, in press; Mehrens, 1992). If further research bears out these initial conclusions, opponents of such assessment programs may argue that there is an intent to discriminate because a less discriminatory alternative (multiple-choice tests) has been replaced with a system that results in greater disadvantage. Employment testing cases have held that cost-effective alternatives with less adverse impact must be considered (*Wards Cove Packing Co. v. Antonio*, 1989). Critics might argue by analogy that such a requirement should also apply to educational assessment. Further discussion of legal and psychometric issues surrounding procedures for minimizing differential item and assessment performance can be found in Phillips (1990, 1993a).

OPPORTUNITY FOR SUCCESS

In addition to protecting the rights of historically disadvantaged groups, the courts have also recognized a more general fundamental fairness requirement that applies to all students. This requirement is part of the substantive due process clause of the Fourteenth Amendment which was described earlier. According to this requirement, assessments must not be arbitrary or capricious and must provide all students with conditions fostering an equal chance for success. Note that this is not a guarantee of equal outcomes but rather of standardized conditions that ensure that no student receives an unfair advantage or penalty. I refer to this concept as *opportunity for success*.

The *Code of Fair Testing Practices* (Joint Committee on Testing Practices, 1988) states:

> Test Developers or Test Users Should: . . . 18. Provide test takers the information they need to be familiar with the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies. *Strive to make such information equally available to all test takers.* [Italics Added] (p. 4)

The *Test Standards* (APA, 1985) recommend:

> [T]est administrators should follow carefully the standardized procedures for administration and scoring specified by the test publisher. Specifications regarding instructions to test takers, time limits, the form of item presentation or response, and test materials or equipment should be strictly observed. (p. 83)

The potential for violation of the opportunity-for-success requirement is of particular concern in the administration of high-stakes performance assessments. The reason is that it is more difficult to maintain standardized assessment administration conditions when students work in groups,, manipulate equipment, or engage in related activities designed to precede the actual assessment.

The following examples illustrate some of these issues. Each example illustrates the potential for unfair advantage or unfair penalties to some students but not others.

Suppose the assessment requires students to measure their classroom and calculate the cost of wall-to-wall carpeting. Suppose further that in most classrooms the students are given yardsticks for measuring, but that one classroom does not have yardsticks available so the teacher gives each student a ruler. Although it is possible to complete the task with a ruler, the task is more difficult and measurement with a ruler may take more time. Measurements made with inefficient measuring instruments are also more prone to error. Thus, failing students who were forced to use a ruler rather than a yardstick could claim unfair treatment because they were required to complete a harder task than the other students. Similar arguments can be made in any situation in which equipment used for a high-stakes assessment task differs from classroom to classroom or school to school.

A related issue involving assessment task equipment is the training students receive in using the equipment. For example, I used to have a calculator that featured reverse-polish notation. By entering a series of operations in reverse order, certain parentheses could be eliminated. Though cumbersome at first, once learned, this feature permitted faster calculation and eliminated errors due to forgetting to enter a last parenthesis. However, facility on this type of calculator made it difficult for me to remember and revert back to the procedures necessary to use a conventional calculator. If I had needed to use a conventional calculator on a high-stakes assessment, I might have become frustrated without some prior practice. This frustration might have negatively impacted my performance. Similarly, any equipment that students may be required to use on an assessment and that is different from what they normally use may pose unfair difficulties due to unfamiliarity and lack of practice.

A possible solution to the above issue would be to allow students to use their own equipment. But this alternative also has standardization problems that might lead to unfairness. For example, suppose some students can afford to purchase programmable calculators while others can afford only a simple, 4-function calculator. Alternatively, a large-scale assessment program might address this issue by furnishing equipment and monitoring training. But in doing so, the program will incur substantial expense. And even then there may be unfairness if some students regularly use calculators they own but others have only one week of practice on calculators furnished specifically for the assessment.

Another opportunity-for-success issue of concern with some performance assessments is the use of cooperative learning. Although cooperative learning may be an effective teaching technique, it may interact in unpredictable ways with the opportunity for success by introducing potential unfairness. For example, suppose a group of four students jointly produces a written essay or completes a science experiment. Suppose further that one academically talented student dominates the process while one weak student contributes little. Should all four students receive the same score? If not, what criteria should be used to judge students' relative contributions to the final product? Should a student who could not have completed the task alone pass the assessment and receive a diploma? What if the best writer in the school went off in a corner and produced an excellent essay without interacting with anyone else? Would that student fail because the process specified by the assessment had not been followed?

Unfairness can also occur when portions of a high-stakes assessment are completed outside of class. For example, a student might be asked to submit a piece of outside writing for evaluation. Will poor students be penalized because there is no parent available to help them with their writing projects? Will some students submit writing products heavily influenced by input from siblings, friends, or tutors? Will some students benefit because their parents provide them access to desktop publishing programs and laser printers that can improve the appearance of the final product and provide illustrative graphics?

Finally, opportunity for success can be denied by differences in a variety of factors under the control of the assessment administrator. Students in different classrooms might be under different time pressures to complete the task due to differences in scheduling, teacher preferences, or other factors. Some administrators who facilitate pretask activities or discussions may cover different information or provide different kinds of assistance in completing the task. Some students may ask questions that result in additional assistance not available to other students. When group work is required, some students may be paired with other students with whom they have difficulty working.

These and a host of other factors may result in assessment experiences that are significantly different for two students with similar knowledge and skills. If one passes and the other does not, the differential assessment conditions may be cited as a significant contributing factor. Although one might believe the latter student would have failed anyway, it may be difficult to convince a judge that all students had an equal opportunity for success on the assessment.

To the extent that specific inequities can be demonstrated, such high-stakes assessments may be judged unfair under the substantive due process standard. To avoid differential opportunities for success in a high-stakes assessment, developers must carefully consider alternative ways to structure the task so that all students respond under equivalent conditions and so that potential inequities are eliminated to the fullest extent possible.

ARTICULATING DEFENSIBLE STANDARDS

Another major issue specifically related to the articulation of goals or standards for high-stakes, large-scale assessments involves the wording of such statements. If students are to be held accountable, goals or standards must specify clearly observable behaviors. One must be careful not to set standards for which it is impossible to reliably determine whether students have met them. For example, an attitude such as appreciation of art or belief in diversity may be fakable if students know what they are expected to say. That is, students may give the "desired answer" while actually believing something entirely different. In such cases, it may be impossible to set up a performance situation that will capture the students' real attitudes.

FREE EXERCISE OF RELIGION AND FREEDOM OF SPEECH

Goals and standards statements may also be subject to First Amendment freedom of speech and free exercise of religion challenges. The First Amendment to the U.S. Constitution provides in part that "Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof." Both the establishment and free exercise clauses of the First Amendment have been used to challenge specific requirements imposed by states and school districts.

In particular, parents with definite religious or political views may object to any goal or standard that appears to require the student to espouse a specific belief or point of view. Performance tasks that require students to read materials about controversial topics and form judgments about alternative actions may trigger such parental challenges. Social studies materials that deal with politically controversial topics may also be targets for challenge.

For example, suppose a reading passage excerpted from a news magazine article described a survey of American religious beliefs in which a random sample of Americans from a variety of denominations responded to questions such as "How often do you attend church services each month?" or "Do you believe in the theory of evolution?" Suppose further that a reading comprehension question following the passage asked which denomination reported attending church services most often each month. Some parents might object to this question as implying a value judgement about the superiority of one denomination over another or as an attempt to influence students for or against their particular denomination.

Similarly, an essay question that asked students to compare and contrast Americans' beliefs about the theory of evolution by religious affiliation might garner the same criticism from parents who object to any attempts by public schools to influence students' religious beliefs. The problems described above could be corrected by substituting a news magazine article describing a survey of the number and types of automobiles owned by a random sample of Americans. Presumably, asking students to summarize information or form judgments about automobile ownership would be much less controversial.

The First Amendment to the U.S. Constitution also provides that "Congress shall make no law . . . abridging the freedom of speech." As long as the speech is not disruptive or vulgar, the courts have held that students are entitled to the free speech rights guaranteed by the First Amendment (*Tinker v. Des Moines Independent Community School Dist.*, 1969; *Bethel School Dist. No. 403 v. Fraser*, 1986). While schools have been allowed some control to maintain student order, complete censorship of the content of student speech is not permitted (*Board of Educ. v. Pico,* 1982; *Hazelwood v. Kuhlmeier,* 1985). Thus, students have been allowed to wear armbands to protest the Vietnam war (*Tinker, 1969*) and to review R-rated movies in the school newspaper (*Desilets v. Clearview Regional Board of Education*, 1994) but not to make campaign speeches containing sexual innuendo (*Bethel School Dist., 1986*). The Supreme Court has also upheld a student's right not to speak (*West Virginia State Bd. of Educ. v. Barnette*, 1943).

Performance standards may be challenged as violating the First Amendment to the extent that they prescribe or proscribe certain viewpoints. For example, the draft physical education standards developed by the National Association for Sport and Physical Education include assessment criteria under a diversity objective that states that students "will not actively voice or exhibit exclusionary behavior" (Diegmueller, 1994). Suppose a school has adopted this standard and a student is found to have violated it because the student advocates separate education for students with severe disabilities.

While such a statement is contrary to the beliefs of those advocating full inclusion, in specific cases it may satisfy the least-restrictive-environment dictates of the Individuals with Disabilities Education Act (1991) and is a position advocated by some parents of disabled students (Schnaiberg, 1994). Requiring the student to support one viewpoint over the other may be seen by some parents as a violation of the student's free speech rights.

The arguments for free speech in schools are similar to those used to strike down hate speech codes (*Doe v. University of Michigan*, 1989). No matter how reprehensible the viewpoint, the Supreme Court has held that the First Amendment protects the speaker from punishment. Withholding a high school diploma for expressing an unpopular viewpoint may be seen as such an inappropriate restriction on free speech.

PARENTAL RIGHTS

Even if the content of multiple-choice questions or performance tasks does not infringe on free speech or religion, parents may still demand the right to preview such assessments prior to administration. A decision is expected soon in a Texas case in which parents have demanded the right to review all state-mandated assessments prior to administration to their children. They argue that denial of assessment review requests violates the constitutional rights of parents explicated in other federal cases.

However, there are no specific parental rights enumerated in the Constitution. The reference to parental rights appears to originate from the Fourteenth Amendment's due process clause together with the Ninth Amendment, which provides that "[t]he enumeration in the Constitution, of certain rights, shall not be construed to deny or disparage others retained by the people." In *Pierce v. Society of Sisters* (1925), the Supreme Court established the right of parents to direct the education and upbringing of their children. Courts have also held that parents have the right to home-school their children, exempt their children from curricula, such as sex education, that may be contrary to their religious beliefs, and obtain special education assistance for qualifying students who attend private schools. In addressing parental rights issues, the Supreme Court has balanced the concerns of the parents against the policy objectives of the state or school district.

With respect to review of assessment questions and tasks, states and school districts have a compelling interest in maintaining the security of the assessment instrument. Because it would be unworkable to attempt to limit review to only those parents with sincerely held religious or political objections, the proposed assessment review would have to be open to all parents. If large numbers of parents were given access to secure assessment instruments, the potential for compromising security would be greatly increased. Elaborate procedures would be necessary to monitor the assessment review to ensure that no copies of the exam were lost and no assessment questions were copied. Despite careful monitoring, some parents might still attempt to memorize assessment questions or content. Performance assessments are particularly vulnerable because they contain fewer questions to remember than multiple-choice exams. If some examinees know the questions or even the topics covered by a performance assessment ahead of time, they will have an unfair advantage over others who do not. Those who argue that this would never happen might recall survey results reported by Popham (1991) in which 57% of California teachers and 60% of California superintendents indicated that practicing on actual test questions from previous forms was appropriate test preparation activity.

It has been suggested that the security concerns described above could be alleviated by review of the assessments after administration with an option to invalidate student answer sheets if parents identify objectionable content. However, such an option would not alleviate the need for security to preserve the technical quality of the assessment program and could adversely affect state and district attempts to evaluate the effectiveness of educational programs. Additionally, school districts would still incur significant expenditures providing adequate facilities, scheduling, and monitoring for parental review sessions.

There are three major reasons why states or districts may need to maintain assessment security after a high-stakes assessment is administered: reuse of items, maintaining equivalent passing standards, and pretesting of items. Thus, compromise of assessment security after administration could still have grave consequences for the quality and viability of an assessment program.

Because tasks and items are expensive to develop, they are often reused in subsequent assessments. The inability to reuse items and the resulting need to develop all new items each year might seriously strain the resources of the assessment program. This could result in a decrease in item quality and could adversely affect the reliability and validity of the assessment instrument.

In addition, some overlap of items from year to year is needed to equate assessment forms so that a consistent passing standard can be maintained. Failure of the equating process could lead to nonequivalent passing standards across consecutive years. Such nonequivalence could result in greater numbers of students failing in a subsequent year because the assessment was more difficult that year. Similarly, students retaking the assessment after a failure might be held to a higher standard the second time. Some educators and parents might view such circumstances as arbitrary, capricious, and unfair.

Finally, states and districts have a compelling interest in pretesting assessment items. Items are pretested by imbedding small numbers of new items among the scored items on an assessment instrument. Students do not know which items count and which are nonsacred pretest items. This technique provides the most accurate data possible for judging item quality because student motivation and assessment conditions are the same for scored and pretested items.

Given the compelling interests of states and districts in maintaining assessment security, the decision in the Texas case could have serious repercussions for high-stakes assessment programs. If the court holds that parents have a right to review assessments administered to their children, considerable expense will be involved in monitoring the review process and in attempting to maintain the technical quality of the assessments. To the extent that assessment data becomes tainted by unfair advantage to some students, the legislative intent of graduation assessments may be subverted. States and school districts will have to make hard decisions about the continued viability and cost effectiveness of assessment programs under such circumstances.

ERRORS OF MEASUREMENT

Performance assessments and their associated standards may be vulnerable to challenge because their errors of measurement tend to be larger than for multiple-choice tests. There are two sources of error of particular concern for performance assessments: generalizability of the sampled tasks and rater effects. Even raters who are well trained to apply previously agreed-upon standards may introduce errors due to irrelevant variables such as handwriting, appearance, speech, and culture. Generalizability of sampled tasks is of concern because a particular small set of tasks may not be a reliable indicator of student achievement of the domain from which they were sampled.

The 0.85 reliability rule of thumb for individual decisions that many professionals have cited in connection with multiple-choice tests may not be appropriate for performance assessments. For performance assessments with few tasks and with scores that are heavily dependent on human judgment of complex interrelated skills, it may be more appropriate to use different measures and possibly lower values. Courts are likely to scrutinize procedures, the performance tasks, and scored student responses in determining what is good enough to withstand challenge.

When considering the reliability of an assessment instrument at the passing standard, it is important to consider the relative seriousness of the two types of error that may occur: false positives (erroneous passers) and false negatives (erroneous failures). The common policy of allowing multiple retakes increases the number of false positives because retested students with true achievement just below the passing standard can eventually expect positive errors of measurement to result in total scores large enough to exceed the passing standard. Multiple retakes also decrease the number of false negatives because the negative errors of measurement responsible for such results will be replaced by a different random and likely more positive error on a subsequent administration. Similarly, when the operational passing standard is set a specified number of standard errors below the judgmental-passing standard to decrease the number of false negatives, the number of false positives will increase. Policymakers may want to consider whether both allowing multiple retakes and lowering the judgmental cut score are necessary to decrease the number of false negatives to an acceptable level and whether the resulting increase in the number of false positives is acceptable.

LEGISLATIVE/ADMINISTRATIVE PASSING STANDARDS

It is common for final decisions on passing standards for statewide assessment programs to be made by legislators, the commissioner of education, and/or the board of education. This is desirable from a legal point of view because the decision is being made by those with the power to pass laws or administrative regulations and to appropriate the necessary funds to administer the assessment program. It is usually the case that the setting of passing standards by such policymakers is informed by standard-setting studies that implemented methodology recommended by measurement professionals. The fact that policymakers may consider additional data when establishing the final standard (e.g., estimated failure rates) does not render their decisions arbitrary. Rather, as the appropriate authority for such decisions, these policymakers are fulfilling their responsibility to consider all the facts and circumstances, including recommendations from standard-setting panels, in formulating the final passing standard.

ASSESSMENT ACCOMMODATIONS FOR THE DISABLED

The Americans with Disabilities Act (ADA, 1990) requires that reasonable accommodations be provided to disabled students. This requirement means that disabled students must be considered when writing goals or standards that apply to all students, when developing assessment items or tasks, and when determining passing or other reporting standards.

PHYSICAL VERSUS COGNITIVE DISABILITIES

Originally, testing accommodations involved compensation for physical disabilities, such as blindness or use of a wheelchair. But recent challenges to assessment practices have focused on accommodations for cognitive disabilities including learning disabilities such as dyslexia, dyscalculia, processing deficits, and so forth. For cognitive disabilities, it is much harder to separate the effects of the disability from the skill being measured.

VALID AND INVALID ACCOMMODATIONS

For assessments and their corresponding score interpretations to be valid, the assessment must measure what is intended, and scores must have the same interpretation for all students. Such is not the case if some students read a reading comprehension test themselves while others have the test read aloud to them. For the latter group, the test is measuring listening comprehension rather than reading comprehension. Similarly, a math computation test for which some students may use calculators and others may not does not measure the same skills for all students. The *Test Standards* (APA, 1985) state:

> . . . unless it has been demonstrated that the psychometric properties of a test, or type of test, are not altered significantly by some modification, the claims made for the test by its author or publisher cannot be generalized to the modified version. (p. 78)

The courts have clearly indicated that reasonable accommodations must compensate for aspects of the disability that are incidental to the skill being measured but that test administrators are not required to change the skill being measured to accommodate a disabled examinee (*Brookhart v. Illinois State Bd. of Educ.*, 1983; *Pandazides v. Virginia Bd. of Educ.*, 1992; *Southeastern Community College v. Davis*, 1979). Drawing the line between a valid and an invalid accommodation requires consideration of the assessment specifications and the uses for the resulting scores. Articulation of an accommodations policy may involve a clash between the goal of inclusion of all students and the validity of the assessment scores for individual, high-stakes decisions.

The most important requirement when balancing inclusion of the disabled against the validity of the assessment is to develop a comprehensive written policy outlining the procedures for requesting accommodations and detailing how decisions will be made regarding specific requests. Both case law and Office for Civil Rights opinion memoranda indicate that accommodation decisions must be made on a case-by-case basis considering the type and degree of disability (*Hawaii State Dept. of Educ.*, 1990).

SCORE NOTATIONS FOR NONSTANDARD ADMINISTRATIONS

There is still some debate about whether score notations for nonstandard administrations may be reported. Privacy statutes preclude reporting the specific disability. However, it may be permissible to report nonstandard administrations when they affect the meaning and interpretation of the assessment score. The *Test Standards* (APA, 1985) summarize the dilemma as follows:

> Many test developers have argued that reporting scores from nonstandard test administrations without special identification (often called "flagging" of test scores) violates professional principles, misleads test users, and perhaps even harms handicapped test takers whose scores do not accurately reflect their abilities. Handicapped people, on the other hand, have generally said that to identify their scores as resulting from nonstandard administrations and in so doing to identify them as handicapped is to deny them the opportunity to compete on the same grounds as nonhandicapped test takers. (p. 78)

The *Code of Fair Testing Practices* (Joint Committee on Testing Practices, 1988) recommends:

> Test Developers Should: . . . 16. . . . Warn test users of potential problems in using standard norms with modified tests or administration procedures that result in non-comparable scores. (p. 3)

Future cases brought under the ADA may provide guidance on this difficult policy issue. In the meantime, policymakers must define "reasonable accommodations" based on all available evidence regarding whether they are incidental to, or confounded with, the skill being measured. If pressured to grant accommodations judged to invalidate assessment score interpretations, policymakers may be able to report such scores separately with a generic designation of "nonstandard administration" after having informed the examinee of the reporting policy and provided the examinee with the option to attempt the assessment without accommodations. However, it is probable that such a policy will be challenged under the ADA. Given the strongly held viewpoints on both sides of the issue, it is currently indeterminate who will prevail.

EXPLICATING ASSUMPTIONS

Another important aspect of the accommodations issue is clear specification of what the assessment is and is not intended to measure. For example, if writing is part of the skill being measured, the standards should say so. If performance in English is expected, this expectation should be clearly communicated. One cannot assume that the courts will accept an unstated assumption that the assessment goals intended the measurement to be in English or that responses must be in writing.

ACCOMMODATION ALTERNATIVES

When there is strong pressure to grant an accommodation that would significantly change the skill being measured, assessment developers may consider changing the nature of the assessment to allow all students access to the accommodation. For example, all students could be allowed to use a calculator on a math assessment that measures concepts and problem solving. Similarly, a reading comprehension assessment could be recast as a communication skills assessment if some students are permitted to have the assessment read to them. Where it is not possible to revise the assessment, but political pressure requires granting an accommodation that changes the nature of the measured skill, assessment developers might consider constructing a separate assessment designed to measure the substituted skill. Alternatively, separate passing standards could be set for an assessment administered in multiple formats. Decisions regarding alternate passing standards and multiple formats should be made based on empirical data where feasible. However, it may not be feasible to collect empirical evidence when the numbers of students with a specific disability is small.

In the long run, allowing all students access to useful accommodations may be fairer to low-achieving students. Research indicates that low achievers tend to be difficult to distinguish from learning disabled students (Clarizio & Phillips, 1989). But because they do not qualify for accommodations, low achievers may not have the same opportunity to demonstrate what they can do. And since many disabilities require additional time, the use of power tests with generous time limits for all students may eliminate potential unfairness.

The arguments presented above are not intended to minimize the inappropriate actions of those who have blatantly and inappropriately discriminated against a disabled person. For example, an employer who rejects a computer programmer in a wheelchair in favor of a less qualified mobile applicant has acted unfairly. Such situations are the kinds of obvious abuses that the federal legislation was designed to correct. The challenge in interpreting the ADA will be to correct the abuses without destroying the interpretability of assessments. Further discussion of valid accommodations-related issues can be found in Phillips (1993a; 1993c; 1994).

ACCOMMODATION ISSUES UNIQUE TO PERFORMANCE ASSESSMENTS

Performance assessments may present some unique accommodations concerns. Some disabled students may have difficulty physically manipulating equipment required to respond to a performance task. In addition, disabilities involving communication deficits may hamper performance across all assessed content areas. This could result in a compounding of the effects of the disability on overall educational achievement.

CONCLUSION

Clearly documented procedures that adhere to professional standards and are administered fairly will be the hallmark of defensible standards. Considering potential legal challenges in the formative stages of an assessment will facilitate the collection and documentation of important evidence.

References

American Psychological Association. (1985). *AERA/APA/NCME Standards for Educational and Psychological Testing*, Washington, D.C.: Author. [Standards For Testing].

Americans with Disabilities Act (ADA, 1990). Pub. L. No. 101-336, 42 U.S.C. § 12101 *et seq.*

Anderson v. Banks, 540 F. Supp. 472 (S.D. Fla. 1981), *reh'g*, 540 F. Supp. 761 (S.D. Ga. 1982).

Beck, M. (in press). "Authentic Assessment" for Large-Scale Accountability Purposes: Balancing the Rhetoric. *Educational Measurement: Issues & Practice.*

Bethel School Dist. No. 403 v. Fraser, 478 U.S. 675, 106 S.Ct. 3159 (1986).

Board of Educ. of Northport-E. Northport v. Ambach, 436 N.Y.S.2d 564 (1981), *aff'd with mod*, 458 N.Y.S.2d 680 (A.D. 1982), *aff'd*, 457 N.E.2d 775 (N.Y. 1983).

Board of Educ. v. Pico, 457 U.S. 853, 102 S.Ct. 2799 (1982).

Clarizio, H.F. & Phillips, S.E. (1989). Defining Severe Discrepancy in the Diagnosis of Learning Disabilities: A Comparison of Methods. *Journal of School Psychology, 27,* 383.

Debra P. v. Turlington, 474 F. Supp. 244 (M.D. Fla. 1979), *aff'd in part, rev'd in part*, 644 F.2d 397 (5th Cir. 1981); *on remand*, 564 F. Supp. 177 (M.D. Fla. 1983) *aff'd*, 730 F.2d 1405 (11th Cir. 1984).

Desilets v. Clearview Regional Board of Education, F. Supp. (1994).

Diegmueller, K. (October 19, 1994). Draft Standards for Health, P.E. Are Released. *Education Week, XIV*(7) 8.

Doe v. University of Michigan, 721 F. Supp. 852 (E.D. Mich. 1989).

Goals 2000: The Educate America Act, Pub. L. No. 103-227 (1994).

Hawaii State Dept. of Educ., 17 EHLR 360 (O.C.R. 1990).

Hazelwood v. Kuhlmeier, 484 U.S. 260, 108 S.Ct. 562 (1985).

Individuals with Disabilities Education Act (I.D.E.A., 1991). Pub. L. No. 102-119, 20 U.S.C. § 1400 *et seq.*

Joint Committee on Testing Practices (1988). Code of Fair Testing Practices in Education, Washington, D.C.: Author.

Mehrens, W.A. (1992). Using Performance Assessment for Accountability Purposes. *Educational Measurement: Issues & Practice, 11*(1) 3.

Nowak, J., Rotunda, R. & Young, J. (1986). *Constitutional Law*. 3rd ed. SS 10.6, 11.4 & 14.1-14.3.

Personnel Administrator v. Feeney, 442 U.S. 256, 99 S.Ct. 2282 (1979).

Pandazides v. Virginia Bd. of Educ., 804 F. Supp. 194 (E.D. Va. 1992).

Phillips, S.E. (1994). High-Stakes Testing Accommodations: Validity Versus Disabled Rights. *Applied Measurement in Education, 7*(2), 93.

Phillips, S.E. (1993a). *Legal Implications of High-Stakes Assessment: What States Should Know.* Oak Brook, IL: North Central Regional Educational Laboratory.

Phillips, S.E. (1993b, March 11, 1993). Legal Issues in Performance Assessment. *Education Law Reporter, 79,* 709.

Phillips, S.E. (1993c, March 25, 1993). Testing Accommodations for Disabled Students. *Education Law Reporter, 80,* 9.

Phillips, S.E. (1991). Diploma Sanction Tests Revisited: New Problems from Old Solutions. *Journal of Law and Education, 20*(2), 175.

Phillips, S.E. (1990, December, 20). The Golden Rule Remedy for Disparate Impact of Standardized Testing: Progress or Regress? *Education Law Reporter, 63,* 383.

Pierce v. Society of Sisters, 268 U.S. 510 (1925).

Pitsch, M. (1994, October 19). Critics Target Goals 2000 in Schools 'War'. *Education Week, XIV*(7), 1.

Popham, J. (1991). Appropriateness of Teachers' Test-Preparation Practices. *Educational Measurement: Issues & Practice, 10*(4), 12.

Schnaiberg, L. (1994, October 5). Chicago Flap Shows Limits of 'Inclusion,' Critics Say. *Education Week, XIV*(5), 1.

Southeastern Community College v. Davis, 442 U.S. 397 (1979).

Tinker v. Des Moines Independent Community School Dist., 393 U.S. 503, 89 S.Ct. 733 (1969).

Village of Arlington Heights v. Metropolitan Housing Dev. Corp., 429 U.S. 252, 97 S.Ct. 555 (1977).

Wards Cove Packing Co. v. Antonio, 109 S.Ct. 211S (1989).

West Virginia State Bd. of Educ. v. Barnette, 319 U.S. 624, 63 S.Ct. 1178 (1943).

BEST COPY AVAILABLE

Legal Defensibility of Standards: Issues and Policy Perspectives

Summary of Break-out Session[1]

The main area of concern in this session was *assessment accommodations*. Two questions were posed: (a) What legal mandates exist that require tests to be administered in special, nonstandard ways to students with identified disabilities? and (b) What effect might these mandates have on setting a common standard of academic achievement, and determining whether students have met it or not? Some participants shared practices in their states, and the discussion focused on practices in those states that use complex performance assessments but must also satisfy the due process requirement of the law. The group members tended to agree that standards can be enforced only when students have had an opportunity to learn what is being tested.

Phillips acknowledged that there is a serious conflict between the efforts to accommodate special needs and the need for fairness, especially with reference to students with cognitive disabilities. She explained that accommodations have meant different things, for example, extending testing time or reading a test aloud. Both practices, it was noted, are in conflict with the principle that all examinees should be administered the same test in the same way in order to provide clear, sound evidence that a standard has been met.

It was emphasized that the issue of accommodations is made more difficult because of the uncertainties in determining whether a student is learning disabled and therefore eligible for accommodation or is just "low achieving" and therefore required to take a test in the standard way. Phillips observed that, generally, more priviledged children whose parents know how to navigate the system, are more likely to be classified as "disabled" whereas poorer children are more likely to be classified as "low achievers." Phillips pointed out that courts will not scrutinize standards or tests until it is disputed that the standards or tests violate an individual's constitutional rights, for example, denying graduation in the case of a student, or denying promotion if the test results are used to make pay and employment decisions. The issue that will be under scrutiny is fairness; the same applies to low-stakes assessments.

Continuing, Phillips explained that there are several issues to be addressed in performance exams: (a) They may require students to write a few essays or prepare a few data charts, but may not have enough items to generalize over a broad range of subjects; (b) the skills involved may be complex and interrelated, and may be difficult and burdensome to document when they have been taught; and (c) there may be difficulties with consistent grading. It was emphasized that changing a test or adjusting a standard undermine the chief purposes of having standards, which are meant to ensure that all students reach a reasonably high level of achievement and that parents and employers can have confidence in school credentials. However, she reiterated that enforcing a standard means making hard decisions and telling some students they do not qualify, which position is in conflict with the policy goals of including all students and minimizing failures and disappointment.

Phillips cautioned that schools with performance standards may face difficulties from judges and administrative officers who focus on helping disabled students. She added that even though a state may have a legitimate measurement argument for its policies, it cannot be certain this argument will prevail.

---

[1] This is a summary of the break-out discussions for Susan Phillips' presentation. The session was facilitated by Ray Fields (National Assessment Governing Board) and recorded by Lawrence Feinberg (National Assessment Governing Board).

It was clarified that the issue of giving exams that enforce *standards in languages other than English* is a policy issue, not a legal one; policymakers must specify that objectives should be met in English. Phillips pointed out that there may be substantial technical problems in applying the same standard on tests in two different languages. However, she indicated that this issue must be considered within the broader policy issue: whether it is better for students in American society to learn English well, or whether learning other skills without English is sufficient.

Issues in Standards Setting:
Making a Decentralized System Responsive to Centralized Policy Making

H. Wesley Smith

Superintendent, Newberg Public Schools, Oregon

ABSTRACT

Standards setting is an effort to make a decentralized public school system responsive to national and state policy decisions. This presentation will explore the hopes, fears, and issues emerging at the local level in the context of state and national standards. In most states, public schools have developed under local control. The result is a tradition-bound, provincial system that is, ostensibly, beyond the influence of central authority. Consequently, as new cultural, social, and economic dynamics demand greater school flexibility and responsiveness to the larger world, the public school system has proven remarkably resistant to change. Policymakers are concerned about how they will find the leadership and resources to overhaul the present system to respond to achievement standards. The professional development initiative required to reach the goal is monumental.

The question is no longer, "Should there be standards-based school reform in American public schools?" It is now, "How shall it develop and be implemented?" Just as Congress profoundly changed the course of special education in the 1970s when it created Public Law 94-142, in March 1994, it redirected the future of American education by enacting Goals 2000: The Educate America Act. The two laws use a similar mechanism to promote change--the creation of standards.

Years of school reform and restructuring efforts have failed to improve student performance, leaving lawmakers and their constituents frustrated. In response, legislators, along with many educators and researchers, have seized upon the idea that setting high standards will improve student performance. Reformers believe that the keys to improving student performance are setting high standards and creating a policy context that aligns curriculum, professional development, and assessments with the standards.

The theory holds that by setting high standards for curricular content and student and school performance, legislators and the public can fashion a mechanism through which educators and the schools may be held accountable. By this mechanism, student learning will improve. The first step in standards-based reform is setting model voluntary national standards. These describe the performance results that students, their parents, and society can expect after students spend 13 years in public schools. States will study, consider, and modify these model standards, then adapt and adopt them. Following adoption, they will design an assessment system aligned precisely with the standards to measure how well students achieve them.

Implementing the standards and assessments will mean that billions of dollars will be spent to prepare curriculum materials that address the standards, to purchase instructional technology appropriate to help students reach the standards, and to train teachers in the methods determined to be most efficient in teaching to the standards. Some of these billions will be federal dollars, but most will be state and local funds. The federal financial participation in school reform will mean only minuscule increases in federal school funding. Yet school reform will become a national effort with a common focus.

Setting standards and assessments is a political process and is the province of policymakers, with the participation of stakeholder groups. By setting standards, policymakers will give schools and teachers direction as to what is to be accomplished and how achievement will be measured. The means of accomplishing the task--how to accomplish it--is left to schools, teachers, parents, and local patrons.

The concept of standards-based reform is simple. Developing and implementing it is complex. It is the complex part that is left to the states, local schools, parents, and teachers.

Those on the receiving end of this new legislation--those who teach students and run the schools--have a few questions and several concerns.

This paper reflects on some of those questions and concerns. Teachers, administrators, parents, board members, and citizens wonder what standard setting will do to and for their schools. This paper identifies issues, phrases questions, conveys opinions, and expresses some of the hopes and fears of school people, parents, and citizens just now undertaking the challenge of implementing standards-based reform.

As awareness of standards-based reform takes hold, local educators and policymakers wonder about the fate of local control and academic freedom. They ask what "high-stakes assessment" really means. How does "top-down support for bottom-up change" work, and can it be trusted? Will high standards become a checklist of minimal competencies? Will devotion to standards narrow the focus of teaching, squeezing human warmth, richness, and vitality out of schools? People wonder and worry about what is meant by the statement, "All students can learn." What are the consequences if all students don't learn?

Questions about standards are more prevalent than opposition to reform, despite the current firestorms of protest by an organized opposition. Most educators and concerned citizens are simply trying to understand the concept and how it will impact the lives of students, their communities, and themselves. It is valuable to consider the thoughts, concerns, and issues on the minds of parents, teachers, and local education leaders. The following pages contain a collection of the issues and challenges that setting standards generates for local practitioners, parents, and communities. These are questions and concerns that should be resolved as standards are established. If standards are to be accepted, supported, and pursued by the American people, these issues must be dealt with early in the process. The urgency of improving student performance has already jeopardized the efficacy of standards-based reform. Assessment development is, in many cases, ahead of standard setting, and is inappropriately driving the system. Both the public and practitioners believe that decisions directly impacting their vital interests are being made before a serious discussion has been initiated.

Astute educators and policymakers understand that standards are powerful tools; they want to be part of ensuring that the tools will be used well.

THE POWER OF STANDARDS

If there is doubt about the power of standards to shape education, consider the historic impact of a single standard, reduced in educational shorthand to three words: "least restrictive environment." This standard requires that special needs students attending public schools that accept federal funds have a right to learn in the least restrictive environment appropriate to their needs and abilities. It revolutionized special education, changed schools, and improved the lives of millions of children. In creating this standard, Congress established an entitlement for certain students and a means by which courts and hearings officers could measure compliance. Courts, because of this federal legislation, assumed an active role in schools. In setting this single standard, Congress engaged the federal judiciary much more actively in school operations.

While it is true that the least-restrictive-environment standard has the force of law and that the Educate America Act (1994) calls only for voluntary national standards, the potential of voluntary national standards is similar to the impact of the least-restrictive-environment standard. The Educate America Act (1994) contains a construction clause that precludes potential litigants from using the act to bring suits based on the voluntary national standards into federal court. States, however, are provided federal incentives to adopt mandatory state standards. Once adopted, such standards may be used to bring litigation in state courts.

By setting standards for educational achievement, will the effect be to define standards for educational practice and thereby, by implication, define malpractice? The question at first glance seems defensive and overly concerned with protecting poor educational practice. However, if mass education is defined through a series of standards, there is real danger of being simplistic in the perception of human development and individuality. An ominous concern about litigation could stifle the vitality and spontaneity of public education.

Justifying setting standards by comparing American schools with the standards of foreign education systems helps little in considering how standards will function in American schools. One viewpoint is that there is no country in the world that has high achievement that does not have high standards. Even if this view point is correct, in this country, the most litigious in the world, explicit standards may have distinctly different consequences than anticipated. The United States operates differently from other nations, and the schools may not respond to centralized standard setting as do schools in other nations. The American legal system is adept at using standards to determine responsibility, rights, and culpability.

Although motivation is hard to ascribe to individuals, let alone to whole legislatures, it appears that lawmakers have seized standards-based reform as the tool with which to make a decentralized public school system respond to centralized policy decisions. It may turn out that the active ingredient in standards-based reform will be the empowerment of the courts to direct school change. Certainly, civil rights and labor legislation has provided opportunities for the courts to direct state and local governments to change policies and practice.

CONTEXT ISSUES

Before Congress considered Goals 2000--even before it considered America 2000--national education standards were being developed. The standards movement was well under way before legislation was conceived. Among contextual issues of concern for many people in the field and among the citizenry are:

1. Neither local educators nor the public are generally aware of the concept of standards-based reform. Allowing standards-setting efforts to get so far ahead of the public and the education community, has jeopardized the entire enterprise of standards-based reform.

2. The process by which standards are being developed is confusing. Content-standards projects do not share a common format or a common definition of a standard, and do not seem to value economy in language or scale.

3. Content standards are being developed by content-centered associations of teachers and professors to the exclusion of other highly interested community and business groups whose involvement is essential to the relevance and acceptance of the standards. Relying on content-centered associations to develop content standards will project the old content-centered, departmentalized school organization plan, which shackles the traditional high school, as the central organizing principle for all schools, kindergarten-grade 12, well into the future. This reliance appears to be a major step into the past.

However vexing these issues may be, they are concerned with decisions previously made and actions taken or under way. There is little to be done about them other than to be aware of the problems they may produce and to work to mitigate their ill effects. Explanations and modifications are in order, but the past cannot be changed. There are more important questions about where things go from here. Over these issues there is control.

PLANNING ISSUES

Senge (1990) warns that today's solutions are the seeds of tomorrow's problems. Care in the process of setting and implementing performance standards may allow that tomorrow's problems are those of society's choosing, not the accidents of oversight. The following thoughts and issues should be considered as standards are set and implemented:

1. It should be recognized that setting standards is a political process. It is better to acknowledge the political nature of the process than to waste time and create controversy by ignoring this fact. Expert assistance in standard setting will be valuable and welcome only to the degree that the experts understand that they are assisting policymakers and not the other way around.

2. The standards set must be credible. They should pass what lawyers call the "laugh test." When citizens and practitioners read a standard and its performance level, they should not dismiss it as frivolous, ridiculously high or low, or foreign to their experience. Unrealistic standards jeopardize the credibility of all the other standards in the system.

3. The number and nature of the standards must be manageable or the standards will be meaningless. If there are scores of standards, they will not be accepted. If standards capture the heart of the nation, there will be fewer standards in 14 or 15 years than there will be in 3 years. States will learn what is important, which standards work, and which don't. Standards will resolve themselves into a short list of consensus all-Americans, common to most states. The standard-setting process should promote and assist this winnowing effect, not work in opposition.

4. Standard setting should constantly consider the process of implementation. Standards should be written so that assessment does not detract from instruction. Neither the standard nor its assessment should inhibit good instruction. Assessment should not absorb undue learning time. Indeed, the standard and its assessment should be structured so that both assist instruction.

5. Terminology must be simplified and clarified. The public, practitioners, and even researchers are confounded by a long list of species of standards. There are content standards, performance standards, delivery standards, school-to-work standards, professional-development standards, and outcome standards. Names give little clue as to how one standard differs from another. Probably, all of the standards mentioned above are outcome standards. They are in bold contrast to the input standards of the past, exemplified by requirements that are the lifeblood of accrediting associations. Finally, there is the hybrid standard, the link between the old input standards and the new output standards--the opportunity-to-learn standard. It appears to be a fuzzy kind of standard, neither output nor input. It implies that should a student or group of students fail to achieve a proper outcome, that failure is an indicator that a proper input was unavailable to the student.

Understandable definitions and simpler, more descriptive language are needed to avoid confusion.

6. The essential attribute of standards-based reform is the alignment of instruction and assessment with standards. Since, according to current management literature, what is measured improves, it is important that assessments measure a desired objective. A measurement that misses its mark contributes to creating a different outcome, one that may not be desired or be of any use.

   Unfortunately, in many schools and states new assessments are being developed before standards have been developed or set. Although this has the effect of setting standards, it is a back-door, unenlightened approach to deciding what students should achieve. It confuses the issue. In designing assessments first, before establishing a standard, there is a tendency to assess those things that are easily tested. In such circumstances, assessment experts play a disproportionately influential role in setting standards. Policymakers, not assessment experts, should lead in determining what schools teach.

7. Standards should not be at odds with current views of instruction. Standards setting should avoid narrowing the field of learning so that it shuts out constructionist views of learning. Among the assumptions of standards-based reform is that "we know how to teach all students"; the way standards are developed, written, and assessed should not deny teachers the very instructional approaches that show the greatest promise for improving the performance of all students.

8. There must be a phase-in plan for standards and their assessment. It must address explaining standards-based reform to educators and the public. In a perfect world, standards would be written and performance levels set through consensus. Once settled upon, assessments would be developed. Before the standards and assessments were put into effect, curriculum materials and tools would be designed and shipped to teachers. The teachers would have already participated in setting the goals and learned their responsibilities and the skills needed to see that students achieved the goals. Then the whole system would, as if responding to the baton of a conductor, commence educating students. The world isn't perfect. States, schools, and teachers are at different stages in reform. Very few students currently nearing high school graduation could meet the performance levels envisioned in the standards under development.

   If practitioners and the American people are to accept content and performance standards, they must trust that today's students will not be penalized for studying under a previous set of expectations. The Kentucky reform experience has taught the value of informing the public of the need to be patient about short-term results. Putting in place competency tests that serve as barriers to students requires a fair period of warning. When tests recently adopted have kept students from graduating, the tests have not survived. To require students to meet new standards without appropriate instruction violates the logic of opportunity-to-learn standards.

   Standards setters are thus presented with a moral dilemma: If the standards set truly represent what young people need to know and be able to do to succeed in society, how can schools graduate them with acknowledged low skill levels that put them at risk? A phase-in plan should at least inform students currently nearing graduation of their apparent deficiencies so that they can, if they choose, remedy their deficiencies.

9. There is a tension between standards and excellence. This tension is producing an apprehension among many educators and parents that standards will describe a competency level that, once achieved, ends a school's commitment to students. Parents of gifted students and educators in schools where budgets are tight are especially worried. It is important that safeguards be established that ensure there are incentives for students, teachers, and schools to excel. These safeguards should include guarantees that teachers and schools may use resources to support the pursuit of excellence by some students, even though a portion of the student body may be failing to reach established standards.

IMPLEMENTATION ISSUES

A Zen proverb admonishes, "Do not confuse the Moon with the finger that points to it."

The concept of setting standards, prescribing performance levels, and creating aligned assessment tools is elegant in its simplicity. It is appealing in its logic. It is utterly useless if ineffectively implemented.

The object of reform is not to impose a standards-based system. The object is to raise the learning and skill levels of all American schoolchildren. Setting standards and assessments is the means, not the end.

To move from theory to implementation is to leave the pristine and enter the messy. Implementation is where the issues are thorniest. Many people are tempted to dwell in theory while neglecting the true task. Researchers and policymakers must remember that it is students, teachers, and others whose lives will be materially changed by the establishment of standards. To achieve true reform, those in the classrooms need freedom to modify theory to what is possible and reasonable in practice.

Even if standards-based reform is exceptionally well accepted and implemented, it will not look or be as theorists have imagined. They must accept and endorse modifications of their plans in local practice.

As the performance standards and assessment are implemented, several additional issues will require resolution:

1. Many people ask, "Who decided that standards-based reform is the way reform must be done?" Many schools have been engrossed in other kinds of school reform over the past decade. Their people believe in what they are doing and do not take kindly to being pulled back from the work of their choosing to follow someone else's star. Despite congressional incentives and mandates in some states, local people interested in education don't believe they have signed on to standards-based reform and are not prepared to surrender what they believe to be the right of local authority to define the nature of their schools.

2. Even among those supporting standards-based reform, many believe that communities should set their own standards, not be tied to national or state standards. It is coming as something of a surprise to some schools and communities that they will no longer be setting their own standards. They do not understand that some authority outside the community will judge the success or failure of their schools. It is unwelcome news to them that the state will tell them what is to be done and that they will be relegated to deciding how to do it and then be accountable to the state.

3. The words "high stakes" are disappearing from the discussions of assessment, just as they disappeared from proposed legislation. Local policymakers take varying interest in, and views of, setting standards, depending on the stakes. Policymakers, sophisticated in reform, are suspicious of the recent omission of the stakes from many of the discussions of standards setting. Some believe that the absence of high stakes negates any promise of improved performance, while others believe that a carrot without a stick has a greater chance for acceptance.

4. In theory, standards allow for centralized policymaking in an otherwise decentralized system and without compromising the decentralized nature of the system. Selling the idea of standards to school boards, site councils, and others accustomed to setting policy and, to a less articulate degree, standards will require a determined effort. It may be of value to convene focus groups of school site council and board members to learn how to deal with this issue.

5. Among the concerns of local policymakers is whether they may set standards higher than those set by the state. May they add standards to those prescribed by the state?

6. There are powerful voices accusing promoters of reform of fiddling while Rome burns. While the education establishment is busying itself with academic standards, the schools have become more dangerous places for students. Poll after poll reports that the public is most concerned about student safety, and justifiably so. Unless education and community leaders address this safety issue, parents, teachers, and citizens will give no credence to the efficacy of learning standards. The establishment and achievement of safety and order standards is a minimum threshold demanded by the public before it commits itself to support other school reform less basic than physical safety.

7. Physical safety is only the most rudimentary of the ingredients necessary for creating opportunities for students to learn. Within opportunity-to-learn standards, there are issues that redefine equity. Old input standards, best illustrated by those enforced by accrediting associations, are obsolete. Opportunity-to-learn standards require that students be given a fair chance to achieve performance standards. The level and equality of opportunity will be judged by results, not input. Given the premise that all students can learn, student failure to achieve the standards may be considered prima facie evidence of a lack of opportunity to learn.

Logic implies that if a child does not learn, the cause for failure must be other than a lack of ability. Among the possible causes may be a lack of application by the student. Yet while Goals 2000 legislation assigns various responsibilities to legislators, school boards, administrators, teachers, parents, and communities for seeing that students learn, it assigns no personal responsibility to students themselves. While the legislation entitles students to the opportunity to learn, it does not assign them a duty to pursue learning.

It is important to consider the responsibilities of various participants in learning when resolving issues surrounding opportunity-to-learn standards. What, if any, are the duties of the state, schools, teachers, parents, and students? What are the remedies if there is a failure to teach or a failure to learn? If a student fails, who is aggrieved--the student, the parents who sent the student to school, the state that paid for the education, the society that supplied the resources and must endure an uneducated citizen, or the teacher whose efforts were in vain?

Until such questions are answered, setting performance standards will create a great many risks for states, educators, and local policymakers.

FINAL COMMENTS

Teachers, principals, school boards, and communities are ready for change. Some are eager; others are grudgingly resigned to reform. Commitment is thin in spots. Local practitioners and the public must be made partners in the reform.

First, teachers and the public must understand what standards-based reform is all about, how it is proposed to work, and what it will mean to their lives. Second, they must believe in the efficacy of standards-based reform. For that to happen, they must be convinced that it is good for the students in their communities.

So far, most of those who will actually implement standards-based reform do not understand it. It has not been well explained to them or to the public. It has yet to be explained in the everyday words that they use. It has not been sold to them. If the standards set do not ring true to teachers and parents, this reform effort will join a long line of reform failures. To succeed, leaders must respond to the issues raised by grassroots participants. A partnership is required between those setting standards and those who must achieve them.

References

Goals 2000: The Educate America Act of 1994, Pub. L. No. 103-227, §1, 108 Stat. 125 (1994).

Senge, P. M. (1990). *The fifth discipline: The art & practice of the learning organization*. New York, NY; Doubleday Currency.

Issues in Standards Setting: Making a Decentralized System Responsive to Centralized Policy Making

Summary of Break-out Session[1]

The group began its discussion with the question: *How should standards be developed and implemented*? The group agreed that there were policy issues at both the state and local levels that were impeding standards-based reform; the group addressed each issue in turn. It was suggested that top-down support for bottom-up solutions was proposed as a viable alternative.

These participants indicated that one issue related to the use of performance standards was how to get beyond the exhortation stage. They pointed out that states initiated the call for standards for the purpose of setting goals and measuring achievement. Federal legislation, including the Goals 2000: Educate America Act and the Elementary and Secondary Education Act (ESEA) reauthorization, attempted to address standards setting; however, the major responsibility for its development and implementation lies at the state and local levels. It was suggested that to accomplish these tasks, it is necessary for all stakeholders to get involved. These stakeholders include the governor, legislature, and agencies at the state level; and schools, parents, and community at the local level.

It was noted that one major hurdle that must be overcome is the *lack of awareness of standards among the general public*, which is also confused with *the definitions of the various types of standards*--content, curriculum, performance, and opportunity-to-learn.

Additionally, it was observed that consideration must be given to *public perception of standards*. The public generally thinks of standards as improvement. However, the public also often regards standards as an imposition. It was recommended that: (a) Standards development should originate at the local level to reflect local concerns; and (b) the public should perceive some degree of local control, i.e., become stakeholders. It was suggested that in this way, there probably would be more widespread agreement, with minor variations, across many local areas, among the standards proposed.

The discussion then focused on *the dynamics of public concern related to education and to standards*. The lead question was: What do people at the local level care about? There was consensus that the general public, which supports schools (primarily through taxes), agrees that there should be standards, but that this term is used to mean that schools and students should do better. For example, graduating seniors should be prepared to become productive members of the workforce. It was indicated that parents, who often make up less than one half of the general public, are more concerned about the "preconditions" of learning--order, discipline, and their children's safety at school in the wake of ongoing violence. It then becomes necessary to raise their comfort level to a point where they can turn their attention to other issues concerning their children. The group agreed that the broader community, whose taxes help support the schools, needs to be made aware of the importance of standards; they do have some idea of standards in that one of their chief concerns is that graduating seniors be prepared to become productive members of the workforce. The group agreed that local citizens should participate in setting standards, at least, in terms of what students should know and be able to do. The participants hypothesized that if local citizens do become involved in setting standards, the standards are unlikely to vary widely from community to community.

---

[1] This is a summary of the break-out discussions for H. Wesley Smith's presentation. The session was facilitated by Daniel Taylor (National Assessment Governing Board) and recorded by Stephen Swearingen (National Assessment Governing Board).

The discussion then focused on *professional development, particularly teacher training and the implementation of standards*. It was noted that as standards have been set and expectations for students have been raised, there has been a reformulation of standards for teachers. However, the following questions were raised: Do teachers understand these standards and can they integrate them into their teaching? It was stated that, too often, teachers are unaware of the standards. For example, the National Council of Teachers of Mathematics (NCTM) has developed standards, but few teachers are familiar with them. On the other hand, teachers themselves say that they just do not have the time. The discussion emphasized that, although teacher education is inherent in much education-related legislation, serious problems exist with staff development and training. There is insufficient distinction between on-the job training for teachers and educational preparation.

The group concluded that strategies need to be developed to reach large numbers of teachers and to facilitate *communication with and among teachers*. Work groups, discipline networks, and newsletters were suggested; the Bay Area Writing Project and a Kentucky monthly newsletter were cited as models for providing training and guidance in both methodology and content.

The group reviewed and critiqued the steps needed to measure progress: (a) develop a baseline, (b) *set targets* that reflect realistic expectations along with detailed time lines, and (c) develop interim indicators. It was noted that targets and time lines may be set for a longer time period than either politicians or the public is willing to accept. The group suggested that the public needs to be given information concerning all the steps in the process of standard setting.

The participants then focused briefly on issues related to *equity and standards for minority groups*. The question posed was: Should standards be lower for disadvantaged students? The discussion focused on the need for all committees with responsibility for the description and development of assessment and standards to be able to anticipate problems of adverse impact related to racial and other minorities. The case of Ohio's 1994 implementation of its graduation standards, which were developed in 1988, was discussed. In this case, the percentage of racial and other minorities failing to meet the standards was greater than that of the group as a whole. One after effect was that students who were not allowed to graduate dropped out of school. The group suggested that additional discussion is needed on this issue.

Program Participants

Phyllis W. Aldrich
Washington-Saratoga-Warren
Hamilton-Essex Counties
WSWHE Board of Cooperative
Educational Services
Donald F. Meyers Center
Henning Road
Saratoga Springs, NY 12866
Telephone: (518) 584-3239
Fax: (518) 583-2780

Ronald A. Berk
Professor
Johns Hopkins University
School of Nursing
Graduate Academic Programs
1830 East Monument Street
Baltimore, MD 21205-2100
Telephone: (410) 955-8211
Fax: (410) 550-5481

Lloyd Bond
Professor
School of Education
University of North Carolina
Greensboro, NC 27412-5001
Telephone: (910) 334-5000
Fax: (910) 334-5882

Robert L. Brennan
E.F. Professor of Educational
Measurement
Director of Iowa Testing Programs
University of Iowa
334 Lindquist Center S.
Iowa City, IA 52242
Telephone: (319) 335-5405
Fax: (319) 335-6038

Belinda L. Collins
Director
Office of Standards Services
National Institute of Standards
and Technology
Building 417, Room 106
Gaithersburg, MD 20899
Telephone: (301) 975-6455
Fax: (301) 963-2871

Joseph Conaty
Acting Director
National Institute on Student
Achievement, Curriculum, and
Assessment
Office of Educational Research
and Improvement
555 New Jersey Avenue, NW
Washington, DC 20208-5573
Telephone: (202) 219-2079
Fax: (202) 219-2030

Emerson Elliott
Commissioner
National Center for Education
Statistics
555 New Jersey Avenue, NW
Washington, DC 20208
Telephone: (202) 219-1761
Fax: (202) 219-1801

Michael Feuer
Director
Board of Testing and Assessment
National Academy of Sciences
2101 Constitution Avenue, NW
Washington, DC 20418
Telephone: (202) 334-3087
Fax: (202) 334-3584

Ronald K. Hambleton
Professor of Education
University of Massachusetts
at Amherst
Hills House South
Room 152
Amherst, MA 01003
Telephone: (413) 545-0262
Fax: (413) 545-4181

Huynh Huynh
Professor
College of Education
University of South Carolina
Columbia, SC 29208
Telephone: (803) 777-7364
Fax: (803) 777-3068
E-Mail:
Huynhhuynh@scarolina.edu

Richard M. Jaeger
Excellence Foundation
Professor and Director
School of Education
University of North Carolina
at Greensboro
Greensboro, NC 27412-5001
Telephone: (910) 334-5000
Fax: (910) 334-5882

Sylvia Johnson
Professor and Editor of the Journal
of Negro Education
Howard University
Room 109
2900 Van Ness Street, NW
Washington, DC 20008
Telephone: (202) 806-8120
Fax: (202) 806-8130

Michael Kane
Professor
Department of Kinesiology
University of Wisconsin at Madison
2000 Observatory Drive
Madison, WI 53706
Telephone: (608) 265-2891
Fax: (608) 262-1656

Wim J. van der Linden
Professor of Educational
Measurement and Data Analysis
Faculty of Educational
Science and Technology
University of Twente
7500 AE Enschede
The Netherlands
Telephone:
011-31-53-893-581
Fax: 011-31-53-356-531
E-Mail:
vanderlinden@edte.utwente.nl

Robert L. Linn
Professor of Education and
Codirector of the National Center
for Research on Evaluation,
Standards, and Student Testing
University of Colorado at Boulder
Campus Box 249
Boulder, CO 80309-0249
Telephone: (303) 492-8280
Fax: (303) 492-7090
E-Mail:
Linnr@spot.Colorado.edu

Samuel A. Livingston
Senior Measurement Statistician
Educational Testing Service
Princeton, NJ 08541
Telephone: (609) 734-1346
Fax: (609) 921-0321

William A. Mehrens
Professor
Michigan State University
462 Erickson Hall
E. Lansing, MI 48824
Telephone: (517) 355-9567
Fax: (517) 353-6393

Samuel Messick
Vice President for Research
Educational Testing Service
Rosedale Road
Princeton, NJ 08540
Telephone: (609) 921-9000
Fax: (609) 734-1090

Pamela A. Moss
Assistant Professor
School of Education
University of Michigan
4220 School of Education
Ann Arbor, MI 48109
Telephone: (313) 747-2461
Fax: (313) 763-1229

Mark D. Musick
President
Southern Regional Education Board
592 Tenth Street, NW
Atlanta, GA 30318-5790
Telephone: (404) 875-9211
Fax: (404) 872-1477

Michael T. Nettles
Professor of Education and
Public Policy
University of Michigan
2035 School of Education Building
610 E. University
Ann Arbor, MI 48109-1259
Telephone: (313) 764-9499
Fax: (313) 764-2510

S.E. Phillips
Professor
Michigan State University
4236 Mariner
Okemos, MI 48864
Telefax: (517) 349-7874

James Popham
Professor Emeritus at the
University of California at
Los Angeles and Director of
IOX Assessment Associates
5301 Beethoven Street, Suite 109
Los Angeles, CA 90066
Telephone: (310) 545-8761
Fax: (310) 822-0269

William T. Randall
Chairman of the National
Assessment Governing Board and
Commissioner of Education
State Department of Education
201 E. Colfax, Street
Denver, CO 80203
Telephone: (303) 866-6806
Fax: (303) 866-6938

Sharon Robinson
Assistant Secretary
Office of Educational Research
and Improvement
U.S. Department of Education
555 New Jersey Avenue NW
Washington, DC 20208
Telephone: (202) 219-1385
Fax: (202) 219-1402

Lorrie A. Shepard
Professor of Education
School of Education
University of Colorado at Boulder
Boulder, CO 80309
Telephone: (303) 492-8108
Fax: (303) 492-7090
E Mail:
Shepardl@spot.Colorado.edu

H. Wesley Smith
Superintendent
Newberg Public Schools
714 East Sixth Street
Newberg, Oregon 97132
Telephone: (503) 537-3211
Fax: (503) 537-9474

Michael J. Zieky
Executive Director
Educational Testing Service
Rosedale Road
Mail Stop 16C
Princeton, NJ 08541
Telephone: (609) 734-5947
Fax: (609) 921-0321

Conference Facilitators

Susan Ahmed
National Center for Education Statistics
555 New Jersey Avenue, NW
Washington, DC 20208-5653
Telephone: (202) 219-1781
Fax: (202) 219-1801

Luz Bay
American College Testing
2201 North Dodge Street
Iowa City, IA 52243
Telephone: (319) 337-1639
Fax: (319) 339-3020

Mary Crovo
National Assessment Governing Board
800 North Capitol Street, NW,
Suite 825
Washington, DC 20002-4233
Telephone: (202) 357-6938
Fax: (202) 357-6945

Ray Fields
National Assessment Governing Board
800 North Capitol Street, NW,
Suite 825
Washington, DC 20002-4233
Telephone: (202) 357-6938
Fax: (202) 357-6945

Jules Goodison
Educational Testing Service
Princeton, NJ 08541
Telephone: (609) 734-5909
Fax: (609) 734-1878

Steven Gorman
National Center for Education Statistics
555 New Jersey Avenue, NW
Washington, DC 20208-5653
Telephone: (202) 219-1937
Fax: (202) 219-1801

Jeanne Griffith
National Center for Education Statistics
555 New Jersey Avenue, NW
Washington, DC 20208-5653
Telephone: (202) 219-1761
Fax: (202) 219-1801

Eugene Johnson
Educational Testing Service
Princeton, NJ 08541
Telephone: (609) 734-5305
Fax: (609) 734-5420

Andrew Kolstad
National Center for Education Statistics
555 New Jersey Avenue, NW
Washington, DC 20208-5653
Telephone: (202) 219-1773
Fax: (202) 219-1801

Stephen Lazer
Educational Testing Service
Princeton, NJ 08541
Telephone: (609) 734-1480
Fax: (609) 734-1878

Susan Loomis
American College Testing
2201 North Dodge Street
Iowa City, IA 52243
Telephone: (319) 337-1048
Fax: (319) 339-3020

Gary Phillips
National Center for Education Statistics
555 New Jersey Avenue, NW
Washington, DC 20208-5653
Telephone: (202) 219-1761
Fax: (202) 219-1801

Mark Reckase
American College Testing
2201 North Dodge Street
Iowa City, IA 52243
Telephone: (319) 337-1105
Fax: (319) 339-3020

Daniel Taylor
National Assessment Governing Board
800 North Capitol Street, NW,
Suite 825
Washington, DC 20002-4233
Telephone: (202) 357-6938
Fax: (202) 357-6945

Paul Williams
Educational Testing Service
Princeton, NJ 08541
Telephone: (609) 734-1427
Fax: (609) 734-1878

Conference Recorders

Salvatore Corrallo
National Center for Education Statistics
555 New Jersey Avenue, NW
Washington, DC 20208-5653
Telephone: (202) 219-1913
Fax: (202) 219-1801

Patricia Dabbs
National Center for Education Statistics
555 New Jersey Avenue, NW
Washington, DC 20208-5653
Telephone: (202) 219-1761
Fax: (202) 219-1801

Larry Feinberg
National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002-4233
Telephone: (202) 357-6938
Fax: (202) 357-6945

Arnold Goldstein
National Center for Education Statistics
555 New Jersey Avenue, NW
Washington, DC 20208-5653
Telephone: (202) 219-1741
Fax: (202) 219-1801

Kristin Keough
National Center for Education Statistics
555 New Jersey Avenue, NW
Washington, DC 20208-5653
Telephone: (202) 219-1680
Fax: (202) 219-1801

Mary Naifeh
National Center for Education Statistics
555 New Jersey Avenue, NW
Washington, DC 20208-5653
Telephone: (202) 219-1551
Fax: (202) 219-1801

Stephen Swearingen
National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002-4233
Telephone: (202) 357-6938
Fax: (202) 357-6945

Sheida White
National Center for Education Statistics
555 New Jersey Avenue, NW
Washington, DC 20208-5653
Telephone: (202) 219-1675
Fax: (202) 219-1801

Shi-Chang Wu
National Center for Education Statistics
555 New Jersey Avenue, NW
Washington, DC 20208-5653
Telephone: (202) 219-1425
Fax: (202) 219-1801

Conference Attendees[1]

Eileen A. Ahearn
Senior Policy Analyst
National Association of State Directors of Special Education
1800 Diagonal Road
Suite 320
Alexandria, VA 22314
Telephone: (703) 519-3800
Fax: (703) 519-3808

Susan Ahmed
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208-5653
Telephone: (202) 219-1781
Fax: (202) 219-1801

Cynthia H. Almeida
Assistant Director
DC Public Schools
Second and Peabody Streets, NW
Washington, DC 20011
Telephone: (202) 576-6288
Fax: (202) 576-7751

David W. Anderson
Georgia Department of Education
Division of Research, Evaluation & Assessment
1870 Twin Towers East
Atlanta, GA 30349
Telephone: (404) 656-2688
Fax: (404) 656-5976

Heibatollah Baghi
Education Associate, Performance Assessment
Delaware Department of Public Instruction
Townsend Building
P.O. Box 1402
Dover, DE 19903-1402
Telephone: (302) 739-2771
Fax: (302) 739-3092

Gaylynn Becker
Department of Public Instruction
State Capitol Building - 11th Floor
600 E. Boulevard
Bismark, ND 58504
Telephone: (701) 224-2755
Fax: (701) 224-2461

Luz Bay
American College Testing
2201 North Dodge Street
Iowa City, IA 52243
Telephone: (319) 337-1639
Fax: (319) 339-3020

Michael Beck
President
BETA, Inc.
35 Guion St.
Pleasantville, NY 10570
Telephone: (914) 769-5235
Fax: (914) 769-4809

Peter Behuniak
Director of Student Assessment
Connecticut Dept. of Education
P.O. Box 2219
Hartford, CT 06145
Telephone: (203) 566-2201
Fax: (203) 566-1625

Isaac Bejar
Educational Testing Service
Rosedale Road - MS 11R
Princeton, NJ 08541
Telephone: (609) 734-5196
Fax: (609) 734-5410

Anita Benson
Project Manager
CTB/McGraw-Hill
20 Ryan Ranch Road
Monterey, CA 93955
Telephone: (408) 393-7813
Fax: (408) 393-7797

Patricia Bent
Education Associate, Assessment
Delaware Department of Public Instruction
P.O. Box 1402
Dover, DE 19903
Telephone: (302) 739-2771
Fax: (302) 737-3092

Sue Betka
Department of Education
FOB-10, Rm. 4251
600 Independence Ave., SW
Washington, DC 20202
Telephone: (202) 401-3939
Fax: (202) 401-0220

Marilyn Binkley
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208-5653
Telephone: (202) 219-1761
Fax: (202) 219-1801

Mary R. Blanton
National Assessment Governing Board Member
Blanton and Blanton
228 West Council Street
Salisbury, NC 28145-2327
Telephone: (704) 637-1100
Fax: (704) 637-1500

George W. Bohrnstedt
Senior Vice President/Office Director
American Institutes for Research
P.O. Box 1113
Palo Alto, CA 94302
Telephone: (415) 493-3550
Fax: (415) 858-0958

---

[1] The list of Conference Attendees is based on conference registration information, as of 10/5/94.

Mary Lyn Bourque
Assistant Director for Psychometrics
National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002-4213
Telephone: (202) 357-6940
Fax: (202) 357-6945

Susan Bowers
Office of Civil Rights
Department of Education
330 C Street, SW, Rm. 5036
Washington, DC 20202
Telephone: (202) 205-8635
Fax: (202) 205-9677

John Braham
Test Consultant
Canadian Nurses Association
50 Driveway
Ottawa, Ontario, Canada K2P 1E2
Telephone: (613) 237-2133
Ext. 214
Fax: (613) 237-3520

Henry I. Braun
Vice President
Educational Testing Service
Mail Stop-19-T, Rosedale Road
Princeton, NJ 08541
Telephone: (609) 734-1239
Fax: (609) 734-5010

Nerissa Bretania-Shafer
Administrator, Research, Planning & Evaluation
Guam Department of Education
P.O. Box DE
Agara, Guam 96910
Telephone: (671) 472-2241
Fax: (671) 477-3407

Benjamin Brown
Director of Accountability Systems
710 James Robertson Parkway
Gateway Plaza, 5th Floor
Nashville, TN 37243-0376
Telephone: (615) 532-4770
Fax: (615) 532-7860

William Brown
Research Professor
UNC Chapel Hill
121 Dunedin Court
Cary, NC 27511
Telephone: (919) 467--2404
Fax: (919) 966-6761

Patricia A. Butler
INTASC Assessment Project
Council of Chief State School Officers
One Massachusetts Ave., NW,
Suite 700
Washington, DC 20001-1431
Telephone: (202) 408-5505
Fax: (202 789-1792

Wayne Camara
Research Scientist
The College Board
45 Columbus Ave.
New York, NY 10023-6992
Telephone: (212) 713-8069
Fax: (212) 713-8181

Dale Carlson
California Department of Education
721 Capitol Mall
Sacramento, CA 95814
Telephone: (916) 657-3011
Fax: (916) 657-4964

Peggy Carr
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1761
Fax: (202) 219-1801

Carmen W. Chapman
Consulting Supervisor
Illinois State Board of Education
100 North First Street
Springfield, IL 62777-0001
Telephone: (217) 782-4823
Fax: (217) 782-6097

Selvin Chin-Chance
Administrator
Test Development Section
Hawaii Department of Education
3430 Leahi Ave., Bldg. D
Honolulu, HI 96815
Telephone: (808) 732-4060
Fax: (808) 732-3701

Rebecca S. Christian
Director, Bureau of Pupil Accountability
Louisiana Department of Education
P.O. Box 94064
Baton Rouge, LA 70804-9064
Telephone: (504) 342-3748
Fax: (504) 342-3684

Brian E. Clauser
Senior Psychometrician
National Board of Medical Examiners
3750 Market St.
Philadelphia, PA 19104
Telephone: (215) 590-9740
Fax: (215) 590-9604

Robert Clemons
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1761
Fax: (202) 219-1801

Allen P. Cook
Director of Research
NCTM
1906 Association Drive
Reston, VA 22091-1593
Telephone: (703) 620-9840
Fax: (703) 476-2970

Salvatore Corrallo
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1913
Fax: (202) 219-1801

John Craig
Assessment Consultant
Illinois State Board of Education
100 North First Street
Springfield, IL 62777-0001
Telephone: (217) 782-4823
Fax: (217) 782-6097

Stephen E. Cramer
Project Coordinator
Test Scoring and Reporting Services
226 Fairfax Hall
University of Georgia
Athens, GA 30602-5589
Telephone: (706) 542-5589
Fax: (706) 542-5364

Elizabeth Creech
Program Director
Georgia Department of Education
1870 Twin Towers East
Atlanta, GA 30334
Telephone: (404) 656-2668
Fax: (404) 656-5976

Marian Crislip
Test Development Specialist
Hawaii Department of Education
3430 Leahi Ave., Bldg. D
Honolulu, HI 96815
Telephone: (808) 735-2237
Fax: (808) 732-3701

Mary Crovo
Assistant Director for Test Development
National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002-4233
Telephone: (202) 357-6941
Fax: (202) 257-6945

Patricia Dabbs
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1933
Fax: (202) 219-1801

Laura K. Damas
Educational Consultant
Ohio Department of Education
65 S. Front St.
Columbus, OH 43215
Telephone: (614) 466-0223
Fax: (614) 728-7434

Charles E. Davis
Executive Director
Statistics and Psychometric Research
Educational Testing Service
Mail Stop 16-T, Rosedale Road
Princeton, NJ 08541
Telephone: (609) 734-5949
Fax: (609) 734-5420

Charles Depascale
Supervisor of Data Analysis
Advanced Systems in Measurement & Evaluation, Inc.
171 Watson Rd.
Dover, NH 03820
Telephone: (603) 749-9102
Fax: (603) 749-6398

Pat DeVito
Director, Office of Outcomes & Assessment
RI Department of Education
22 Hayes St.
Providence, RI 02908
Telephone: (401) 277-3126
Fax: (401) 277-4979

Sean Donovan
Measurement Statistician
Kentucky Dept. of Education
OCAA - 500 Mero St.
Frankfort, KY 40601
Telephone: (502) 564-4394
Fax: (502) 564-7749

Gordon Ensign
Supervisor, Curriculum & Assessment
State Superintendent of Public Instruction
P.O. Box 47200
Old Capitol Bldg.
Olympia, WA 98504
Telephone: (360) 753-3449
Fax: (360) 586-2728

Rajah K. Farah
Assessment Consultant
Wisconsin Department of Public Instruction
125 S. Webster Street
Madison, WI 53707-7841
Telephone: (408) 267-9283
Fax: (408) 266-8770

Larry Feinberg
Assistant Director for Reporting & Dissemination
National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002-4233
Telephone: (202) 357-6942
Fax: (202) 357-6945

Steve Ferrara
State Director of Student Assessment
Maryland Department of Education
PRIM/Assessment
200 W. Baltimore St.
Baltimore, MD 21201
Telephone: (410) 767-0081
Fax: (410) 333-0052

Ray Fields
Assistant Director for Policy & Research
National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002-4233
Telephone: (202) 357-6938
Fax: (202) 357-6945

Thomas H. Fisher
Administrator
Department of Education
325 W Gaines Street, FEC 701
Tallahassee, FL 32399-0400
Telephone: (904) 488-8198
Fax: (904) 487-1889

Chester E. Finn. Jr.
National Assessment Governing Board Member
John M. Olin Fellow
Hudson Institute
1015 18th Street, NW
Suite 200
Washington, DC 20036
Telephone: (202) 223-5450
Fax: (202) 223-9226

Valeria A. Ford
Director of Student Assessment
District of Columbia Public Schools
Rabaut Administration Building
Room 113
2nd and Peabody Sts., NW
Washington, DC 20011
Telephone: (202) 576-6288
Fax: (202) 576-7751

Robert Forsyth
University of Iowa
320 LC
Iowa City, IA 52242
Telephone: (319) 335-5412
Fax: (319) 335-6038

Shari L. Francis
Director of State Relations
National Council for Accreditation of Teacher Education
2010 Massachusetts Ave., NW
Washington, DC 20036
Telephone: (202) 466-7496
Fax: (202) 296-6620

Chester W. Freed
Delaware Dept. of Public
Instruction
Townsend Building
Dover, DE 19901
Telephone: (302) 739-4583
Fax: (302) 739-4221

James Friedebach
Director of Assessment
205 Jefferson City
Jefferson City, MO 65101
Telephone: (314) 751-3545
Fax (314) 751-9434

Edward Fuentes
U.S. Department of Education
Office of Educational Research & Improvement
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1761
Fax: (202) 219-1801

Carol Fuller
Assistant Executive Director
National Institute of Independent Colleges & Universities
122 C St., NW, Suite 750
Washington, DC 20001
Telephone: (202) 347-7512
Fax: (202) 628-2513

Larry Gabbert
Education Associate
Delaware Department of Public Instruction
P.O. Box 1402
Dover, DE 19903
Telephone: (302) 767-0368
Fax: (302) 566-3867

Robert Gabrys
Assistant State Superintendent for Research & Development
MD State Department of Education
200 West Baltimore Street
Baltimore, MD 21201
Telephone: (410) 767-0368
Fax: (410) 333-3867

Matthew Gandal
Research Associate
American Federation of Teachers
555 New Jersey Ave, NW
Washington, DC 20001
Telephone: (202) 879-4501
Fax: (202) 879-4537

Kathryn Gang
Director of Education Programs
Research Development Corporation
2875 Towerview Road, Suite A-4
Herndon, VA 22071
Telephone: (703) 904-1808
Fax: (703) 904-1812

Helen Ganson
Scottish Examination Board
Ironmills Road
Dalkeith, Scotland
EH 221LE
Telephone: (031) 663-6601
Fax: (031) 654-2664

Jack Gilsdorf
Director, Assessment and Evaluation
Nebraska Department of Education
301 Centennial Mall South
Lincoln, NE 68509
Telephone: (402) 471-2444
Fax: (402) 471-0117

Arnold Goldstein
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1761
Fax: (202) 219-1801

Jules Goodison
Deputy Director
Educational Testing Service
Rosedale Rd.
Princeton, NJ 08541
Telephone: (609) 734-5909
Fax: (609) 734-1878

Steven Gorman
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1937
Fax: (202) 219-1801

B.J. Granbery
Chapter 1, Director (Title I)
Montana Office of Public Instruction
State Capitol
P.O. Box 202501
Helena, MT 59620-2501
Telephone: (406) 444-4420
Fax: (406)444-3924

George Gray
American College Testing
2201 N. Dodge St.,
P.O. Box 168
Iowa City, IA 52243
Telephone: (319) 337-1168
Fax: (319) 339-3021

Bert F. Green, Jr.
Professor
Johns Hopkins University
Psychology Dept., J.H.U.
3400 Charles St.
Baltimore, MD 21218
Telephone: (410) 516-7074
Fax: (410) 516-4478

Donald Ross Green
Chief Research Psychologist
CTB/McGraw Hill
20 Ryan Ranch Road
Monterey, CA 93940
Telephone: (408) 393-0700
Fax: (408) 393-7825

Jeanne Griffith
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1676
Fax: (202) 219-1801

Michael J. Guerra
National Assessment Governing Board Member
Executive Director
NCEA, Secondary School Dept.
1077 30th Street, NW
Suite 100
Washington, DC 20007
Telephone: (202) 337-6232
Fax: (202) 333-6706

Marilyn Hala
Director of Professional Programs
National Council of Teachers of Mathematics
1906 Association Drive
Reston, VA 22091
Telephone: (703) 620-9840
Ext.151
Fax: (703) 476-2970

Carroll Hall
Director of Assessment & Evaluation
New Mexico Department of Education
Capitol Complex
Santa Fe, NM 87501
Telephone: (505) 827-6524
Fax: (505) 827-6696

Ron Hall
Department of Education
Office of the Commissioner
555 New Jersey Ave., NW
Room 400
Washington, DC 20208
Telephone: (202) 219-1839
Fax: (202) 219-1736

Linda Hansche
Director, Assessment & Technology
Georgia State University Resources
P.O. Box 858
University Plaza
Atlanta, GA 30303
Telephone: (404) 651-4230
Fax: (404) 651-4226

Kelley Henderson
Policy Specialist for Governmental Relations
Council for Exceptional Children
1920 Association Drive
Reston, VA 22091-1589
Telephone: (703) 264-9418
Fax: (703) 620-4334

James F. Hertzog
Chief/Division of Evaluation & Reports
PA Department of Education
333 Market Street
Harrisburg, PA 11126-0333
Telephone: (717) 787-4234
Fax: (717) 783-6642

E. Lynn Holley-Green
Development Manager
CTB/McGraw-Hill
20 Ryan Ranch Road
Monterey, CA 93940
Telephone: (408) 393-7750
Fax: (408) 393-7607

Eugene Johnson
Director, Research
Educational Testing Service
Princeton, NJ 08541
Telephone: (609) 734-5305
Fax: (609) 734-5420

Janet Johnson
Educational Testing Service
Princeton, NJ 08541
Telephone: (609) 734-5598
Fax: (609) 734-5420

Lyle V. Jones
Research Professor
Univ. of North Carolina at Chapel Hill
CB 3270, UNC-CH
Chapel Hill, NC 27599
Telephone: (919) 962-2325
Fax: (919) 962-2537

Ann Jungeblut
Senior Research Associate
Educational Testing Service
Rosedale Road
Princeton, NJ 08541
Telephone: (609) 734-1090
Fax: (609) 734-5420

Stuart R. Kahl
Vice President
Advanced Systems in Measurement & Evaluation, Inc.
171 Watson Rd.
Dover, NH 03820
Telephone: (603) 749-9102
Fax: (603) 749-6398

BEST COPY AVAILABLE

Barbara A. Kapinus
Senior Program Coordinator
Council of Chief State
School Officers
One Massachusetts
Avenue, NW, Suite 700
Washington, DC 20001-1431
Telephone: (202) 336-7058
Fax: (202) 789-5305

James Karon
Coordinator/State Assessment
Program
Rhode Island Department of
Education
22 Hayes St.
Providence, RI 02908
Telephone: (401) 277-3126
Fax: (401) 277-4979

Kristin Keough
National Center for Education
Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1680
Fax: (202) 219-1801

Mary M. Kino
Project Director, Psychometrics &
Tech. Application
The Psychological Corporation
555 Academic Court
San Antonio, TX 78204
Telephone: (210) 299-3670
Fax: (210) 270-0327

Irwin Kirsch
Director Research
Educational Testing Service
Rosedale Road
Princeton, NJ 08541
Telephone: (609) 734-1516
Fax: (609) 734-5420

Andrew Kolstad
National Center for Education
Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1773
Fax: (202) 219-1801

Nancy Krause
Director
Los Angeles County Office of
Education
9300 Imperial Hwy.,
Room 210
Downey, CA 90242
Telephone: (310) 922-6381
Fax: (310) 922-6299

Samuel E. Krug
MetriTech, Inc.
111 N. Market Street
Champaign, IL 61820
Telephone: (217) 398-4868
Fax: (217) 398-5798

David Kysilko
Director of Publications
National Assoc. of State Boards of
Education (NASBE)
1012 Cameron St.
Alexandria, VA 22314
Telephone: (703) 684-4000
Fax: (703) 836-2313

Millie Lanauze
U.S. Department of Education
OBEMLA
330 C Street, SW
Room 5616
Washington, DC 20202
Telephone: (202) 205-9475
Fax: (202) 205-8737

Archie E. Lapointe
Executive Director
Educational Testing Service
Rosedale Road
Princeton, NJ 08541-6710
Telephone: (609) 734-5890
Fax: (609) 734-1878

Stephen Lazer
Program Administrator
Educational Testing Service
Rosedale Road
Princeton, NJ 08541
Telephone: (609) 921-9000
Fax: (609) 734-1480

Zoe E. Leimgruebler
Director of Student Assessment
Oklahoma State Department of
Education
2500 N. Lincoln Boulevard
Suite 2-16
Oklahoma City, OK 73105
Telephone: (405) 521-3341
Fax: (405) 521-6205

Mary Lindquist
Past President
National Council of Teachers of
Mathematics
14 Seventh Street
Columbus, GA 31901
Telephone: (706) 323-7520
Fax: (706) 569-3134

David Lohman
Professor
366 Lindquist Center
University of Iowa
Iowa City, IA 52242
Telephone: (319) 335-8229
Fax: (319) 335-5386

Susan Cooper Loomis
NAEP ALS Project Director
American College Testing
2201 N. Dodge Street
Iowa City, IA 52243
Telephone: (319) 337-1048
Fax: (319) 339-3021

Brenda Loyd
University of Virginia
272 Ruffner Hall, UVA
405 Emmet St.
Charlottesville, VA 22901
Telephone: (804) 924-0824
Fax: (804) 924-0747

Richard M. Luecht
Senior Psychometrician
National Board of Medical
Examiners
3750 Market Street
Philadelphia, PA 19104
Telephone: (215) 590-9646
Fax: (215) 590-9604

Gail S. MacColl
Evaluator
U.S. General Accounting Office
Room 5853
441 G St., NW
Washington, DC 20548
Telephone: (202) 512-5108
Fax: (202) 512-2622

Duncan MacQuarrie
Supervisor, Research & Evaluation
State Superintendent of Public Instruction
P.O. Box 47200
Old Capitol Bldg.
Olympia, WA 98504
Telephone: (206) 753-3449
Fax: (206) 586-2728

Barbara Marenus
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1761
Fax: (202) 219-1801

Wayne H. Martin
Assessment Director
Colorado Department of Education
201 East Colfax Avenue
Room 501
Denver, CO 80203
Telephone: (303) 866-6850
Fax: (303) 830-0793

John Martois
Consultant
Los Angeles County Office of Education
9300 Imperial Hwy.,
Room 210
Downey, CA 90242
Telephone: (310) 922-6304
Fax: (310) 922-6299

Rodney A. McCloy
Senior Scientist
Human Resources Research Organization
66 Canal Center Plaza, Suite 400
Alexandria, VA 22314
Telephone: (703) 706-5653
Fax: (703) 548-5574

Sally N. McConnell
Director of Status & Advocacy
National Association of Elementary School Principals
1615 Duke St.
Alexandria, VA 22314
Telephone: (703) 684-3345
Fax: (703) 5448-6021

Don McLaughlin
Chief Scientist
American Institutes for Research
1791 Arastradero Road
P.O. Box 1113
Palo Alto, CA 94302
Telephone: (415) 493-3550
Fax: (415) 858-0958

James E. McLean
Research Professor and Assistant Dean
The University of Alabama
P.O. Box 870231
Tuscaloosa, AL 35487-0231
Telephone: (205) 348-6874
Fax: (205) 348-6873

William McMillan
Director of Assessment
MN State Department of Education
550 Cedar Street
St. Paul, MN 55101
Telephone: (612) 296-6002
Fax: (612) 296-3348

Cheryl Mercer
Education Program Consultant
Kansas State Board of Education
120 S.E. 10th St.
Topeka, KS 66610
Telephone: (913) 296-3996
Fax: (913) 296-7933

Hillary Michaels
Maryland State Department of Education
200 W. Baltimore St.
Baltimore, MD 21201-2595
Telephone: (410) 333-8431
Fax: (410) 333-2017

Jason Millman
National Assessment Governing Board Member
Cornell University
405 Kennedy Hall
Ithaca, NY 14853-4203
Telephone: (607) 255-7704
Fax: (607) 255-7905

Carol Morrison
Psychometrician
National Board of Medical Examiners
3750 Market St.
Philadelphia, PA 19104
Telephone: (215) 590-9745
Fax: (215) 590-9604

Grisel Munoz
Director of Evaluation
General Council on Education
P.O. Box 5429
Hato Rey, Puerto Rico 00919
Telephone: (809) 764-0910
Fax: (809) 764-0820

Mary Naifeh
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1551
Fax: (202) 219-1801

Patricia A. Nathan
Testing Director
Department of Education
St. Thomas
#44-16 Kongens Gade
St. Thomas, VI 00801
Telephone: (809) 779-7121

Ken Nelson
National Education Goals Panel
1850 M Street, NW
Suite 270
Washington, DC 20036
Telephone: (202) 632-0952
Fax: (202) 632-0957

Dori Nielson
Senior Education Analyst
Montana Office of Public Instruction
State Capitol
P.O. Box 2501
Helena, MT 59620-2501
Telephone: (406) 444-3656
Fax: (406) 444-3924

Christine O'Sullivan
Science Coordinator
Educational Testing Service
Mail Stop; 30-E,
P.O. Box 6710
Princeton, NJ 08541-6710
Telephone: (609) 734-1918
Fax: (609) 734-1878

Laurence Ogle
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1761
Fax: (202) 219-1801

John F. Olsen
NAEP Program Director
Educational Testing Service
P.O. Box 6710
Princeton, NJ 08541-6710
Telephone: (609) 734-1987
Fax: (609) 734-1878

Eugene Owen
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1761
Fax: (202) 219-1801

Tej Pandey
Administrator, Research, Evaluation & Technology
California Dept. of Education
721 Capitol Mall, 4th Floor
Sacramento, CA 95814
Telephone: (916) 657-4301
Fax: (916) 657-4978

Eugene T. Paslov
Executive Director
New Standards Project
700 11th St., NW
Suite 750
Washington, DC 20001
Telephone: (202) 783-3668
Fax: (202) 783-3672

Lois Peak
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1761
Fax: (202) 219-1801

Samuel Peng
Department of Education
NCES/SSMD
Room 408F
Washington, DC 20208
Telephone: (202) 219-1643
Fax: (202) 291-1801

Gary Phillips
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1761
Fax: (202) 219-1801

Barbara Wells Price
Manager
Performance Assessment Scoring
National Computer Systems
1070 William Street
Iowa City, IA 52240
Telephone: (319) 339-6409
Fax: (319) 339-6593

Cynthia Prince
National Education Goals Panel
1255 22nd Street, NW
Suite 502
Washington, DC 20037
Telephone: (202) 632-0952
Fax: (202) 632-0957

Elizabeth Rangeo
Director, Learning Research and Development Center
University of Pittsburgh
833 LRDC, 3939 O'Hara St.
Pittsburgh, PA 15260
Telephone: (412) 624-7450
Fax: (412) 624-9149

Mark D. Reckase
Assistant Vice President
American College Testing
2201 N. Dodge St.
Iowa City, IA 52243
Telephone: (319) 337-1105
Fax: (319) 339-3021

Rod Riffel
Senior Policy Analysis
National Education Association
1201 16th Street, NW
Washington, DC 20036
Telephone: (202) 822-7384
Fax: (202) 822-7482

Douglas A. Rindone
Chief, Bureau of Evaluation & Student Assessment
CT State Department of Education
P.O. Box 2219
165 Capitol Ave.
Hartford, CT 06145
Telephone: (203) 566-1684
Fax: (203) 566-1625

Janelle Rivers
Education Associate
SC Department of Education
1429 Senate Street
Room 607
Columbia, SC 29201
Telephone: (803) 734-8266
Fax: (803) 734-8264

Jeff Rodamar
Department of Education
Planning and Evaluation Services
600 Independence Ave., SW
Room 4136
Washington, DC 20202
Telephone: (202) 205-5046
Fax: (202) 401-3036

Edward D. Roeber
Director, Student Assessment Programs
Council of Chief State School Officers
1 Massachusetts Ave., NW
Suite 700
Washington, DC 20001
Telephone: (202) 336-7045
Beeper: (202) 789-1792
Fax: (202) 408-8072

Edgar D. Ross
National Assessment Governing Board Member
Territorial Court of the Virgin Isl.
Judge
RFO2 Box 9000
Christiansted, St. Croix
U.S. Virgin Islands, 00821
Telephone: (809) 778-9750

Shelley Loving-Ryder
VA Department of Education
P.O. Box 2120
Richmond, VA 23216-2120
Telephone: (804) 225-2102
Fax: (804) 371-8978

Paul Sandifer
Assistant Vice President
American College Testing
2201 North Dodge Street
P.O. Box 168
Iowa City, IA 52243
Telephone: (319) 339-3079
Fax: (319) 339-3021

Joan Seamon
Senior Associate
National Assn. of State Boards of Education
1012 Cameron St.
Alexandria, VA 22314
Telephone: (703) 684-4000
Fax: (703) 836-2313

Alex Sedlacek
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1761
Fax: (202) 219-1801

Ramsay W. Selden
Director, State Education Assessment Center
CCSSO
One Massachusetts Ave., NW
Suite 700
Washington, DC 20001-1431
Telephone: (202) 336-7010
Fax: (202) 789-1792

Juliet P. Shaffer
Senior Lecturer
Dept. of Statistics, Univ. of California
367 Evans Hall
Berkeley, CA 94720-3860
Telephone: (510) 549-3596
Fax: (510) 642-7892

Sharif Shakrani
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1761
Fax: (202) 219-1801

Stephen G. Sireci
Senior Psychometrician
American Council on Education
One Dupont Circle, NW
Washington, DC 20036-1163
Telephone: (202) 939-9490
Fax: (202) 775-8578

Ernest N. Skakun
Director of Psychometrics
University of Alberta, Faculty of Medicine
Division of Studies in Medical Education
25300 Walter Mackenzie Centre
Edmonston, Alberta
Canada T6G 2R7
Telephone: (403) 492-6776
Fax: (403) 492-5487

Janice Smith-Dann
Program Specialist
FL Department of Education
Florida Education Center
325 W. Gaines Street, Suite 701
Tallahassee, FL 32399
Telephone: (904) 488-8198
Fax: (904) 487-1889

Diane L. Smolen
Supervisor, Michigan Educational Assessment Program
Michigan Department of Education
P.O. Box 30008
Lansing, MI 48909
Telephone: (517) 373-8393
Fax: (517) 335-1186

Fran Stancavage
Senior Research Scientist
American Institutes for Research
P.O. Box 1113
Palo Alto, CA 94302
Telephone: (415) 493-3550
Fax: (415) 858-0958

Donald Stewart
President
College Board
45 Columbus Avenue
New York, NY 10023
Telephone: (212) 713-8036
Fax: (212) 713-8282

Rajah G. Subhiyah
Senior Psychometrician
National Board of Medical Examiners
3750 Market Street
Philadelphia, PA 19104
Telephone: (215) 590-9741
Fax: (215) 590-9453

Stephen Swearingen
Budget Officer
National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002-4233
Telephone: (202) 357-6938
Fax: (202) 357-6945

David Sweet
Senior Researcher
Office of Educational Research & Improvement
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1748
Fax: (202) 219-2030

Robert Sykes
Research Scientist
CTB/McGraw-Hill
20 Ryan Ranch Rd.
Monterey, CA 93940
Telephone: (408) 393-0700
Fax: (408) 393-7825

Daniel Taylor
Deputy Director
National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002-4233
Telephone: (202) 357-6944
Fax: (202) 357-6945

Mary Anne Tharin
Staff Consultant
NC Educational Standards & Accountability Commission
116 W. Jones St.
Raleigh, NC 27603
Telephone: (919) 715-0893
Fax: (919) 733-1753

Cheryl Tibbals
Director of State and Local Relations
New Standards Project
700 11th St., NW
Suite 750
Washington, DC 20001
Telephone: (202) 783-3668
Fax: (202) 783-3672

Roy Truby
Executive Director
National Assessment Governing Board
800 N. Capitol Street, NW
Suite 825
Washington, DC 20002-4233
Telephone: (202) 357-6938
Fax: (202) 357-6945

Gloria Turner
Coordinator of Student Assessment
Alabama State Department of Education
50 North Ripley St.,
Room 3304
Montgomery, AL 36130-3901
Telephone: (205) 242-8038
Fax: (205) 242-0482

Jon Twing Jr.
Project Director
The Psychological Corp.
555 Academic Court
San Antonio, TX 78204
Telephone: (210) 299-2754
Fax: (210) 270-0327

Gordon Wainwright
Research Scientist
CTB/McGraw-Hill
20 Ryan Ranch Rd.
Monterey, CA 93940
Telephone: (408) 393-0700
Fax: (408) 393-7825

Janet E. Wall
Manager, ASVAB Career Exploration Program
Defense Manpower Data Center
99 Pacific St., Suite 115A
Monterey, CA 93940
Telephone: (408) 655-0400
Fax: (408) 656-2087

Carolyn Warren
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1761
Fax: (202) 219-1801

Don Watson
Senior Consultant, Student Assessment
Colorado Department of Education
201 E. Colton, Room 501
Denver, CO 80111
Telephone: (303) 866-8854
Fax: (303) 830-0793

Joe Webb
Staff Consultant
NC Education Standards and Accountability Commission
116 W. Jones St.
Raleigh, NC 27603
Telephone: (919) 715-0893
Fax: (919) 733-1753

Mel Webb
Acting Director
Assessment Office
Philadelphia School District
Room 403
21st Street & The Parkway
Philadelphia, PA 19103-1099
Telephone: (215) 299-2525 or
Telephone: (215) 299-7758
Fax: (215) 299-3408

Ann S. White
Assistant Director
Professional Assessment Services
American College Testing
2201 N. Dodge St.,
P.O. Box 168
Iowa City, IA 52243
Telephone: (319) 337-1125
Fax: (319) 339-3020

Kathleen White
Senior Evaluator
GAO
441 G St., NW, Rm. 5853
Washington, DC 20548
Telephone: (202) 512-8512
Fax: (202) 512-2622

Sheida White
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1675
Fax: (202) 219-1801

Paul L. Williams
Executive Director
Center for Assessment
Education Testing Center
Rosedale Road, Box 6710
Princeton, NJ 08541-6710
Telephone: (609) 734-1427
Fax: (609) 734-5410

Mary Ann Wilmer
Executive Officer
National Assessment Governing Board
800 N. Capitol St., NW
Suite 825
Washington, DC 20002-4233
Telephone: (202) 357-6906
Fax: (202) 357-6945

Lauress Wise
Defense Manpower Data Center
99 Pacific Street
Suite 155A
Monterey, CA 93940-2453
Telephone: (408) 655-0440
Ext. 4271
Fax: (408) 656-2087

Shi-Chang Wu
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208
Telephone: (202) 219-1425
Fax: (202) 219-1801

Emily Wurtz
National Education Goals Panel
1850 M Street, NW
Suite 270
Washington, DC 20036
Telephone: (202) 632-0952
Fax: (202) 632-0957

Elaine Young
State NAEP Coordinator
New Jersey Department of Education
240 West State St.
Trenton, NJ 08625
Telephone: (609) 777-3671
Fax: (609) 984-6032

Michael Young
Research Scientist
CTB/McGraw-Hill
20 Ryan Ranch Rd.
Monterey, CA 93940
Telephone: (408) 393-7292
Fax: (408) 373-7825

Gloria Zyskowski
2201 Donley Drive
Suite 100
Austin, TX 78758
Telephone: (512) 835-4833
Fax: (512) 835-8083